Εθνικό Μετσόβιο Πολυτεχνείο
Σχολή Εφαρμοσμένων Μαθηματικών και Φυσικών Επιστημών
Τομέας Μαθηματικών

Στατιστική Ανάλυση Δεδομένων Εργατικού Δυναμικού με χρήση της R

Διπλωματική Εργασία

Μαρία Ντάβου

Επιβλέπων: Δημήτριος Φουσκάκης
Καθηγητής Ε.Μ.Π.

Περίληψη

Η παρούσα διπλωματική εργασία ασχολείται με την μελέτη και ανάλυση των δεδομένων εργατικού δυναμικού, που προέρχονται από την αντίστοιχη έρευνα της Ελληνικής Στατιστικής Αρχής, η οποία παρουσιάζεται ενδελεχώς στο πρώτο κεφάλαιο. Η κύρια ανάλυση βασίζεται στα δεδομένα που συλλέχθησαν για το έτος 2022, αλλά χρησιμοποιούνται και δεδομένα των ετών 2018 έως 2022.

Αρχικά, πραγματοποιείται Περιγραφική Στατιστική με σκοπό τον έλεγχο της κατάστασης απασχόλησης στην Ελλάδα. Η μελέτη έγινε σε κατηγορίες που αφορούν το σύνολο του εργαστικού δυναμικού, όπως η ηλικία, η εκπαίδευση, η Περιφέρεια, η υπηκοότητα και ο βαθμός αστικοποίησης. Στην συνέχεια, διαχωρίζουμε την ανάλυση σε άτομα που εργάζονται και σε ανέργους, χρησιμοποιώντας διαφορετικές κατηγορίες για την κάθε κατάσταση απασχόλησης. Τέλος, πραγματοποιείται συγκριτική ανάλυση σε κατηγορίες που αφορούν την ανεργία για τα έτη 2018 έως 2022, με σκοπό να αντιληφθούμε πως μεταβάλλεται το φαινόμενο της ανεργίας. Για την Περιγραφική Στατιστική χρησιμοποιήθηκαν τα πακέτα ggplot2 και data.table της R.

Στην συνέχεια της διπλωματικής, προσαρμόζεται και παρουσιάζεται το μοντέλο της Λογιστικής Παλινδρόμησης για τα εν λόγω δεδομένα, όπου μελετάται η εξάρτηση της ανεργίας από τους παράγοντες της ηλικίας, του φύλου, της εκπαίδευσης και της υπηκόοτητας (αφού πρώτα έχει παρουσιαστεί το θεωρητικό πλαίσιο της Λογιστικής Παλινδρόμησης). Τέλος, ελέγχεται μέσω της Ταξινόμησης και της Διασταυρωμένης Επικύρωσης η ικανότητα του μοντέλου να προσαρμοστεί ορθά σε μελλοντικά δεδομένα.

Επιπλέον, στο τελευταίο κεφάλαιο παρουσιάζεται ο κώδικας των διαγραμμάτων και των πινάκων που δημιουργήθηκαν με το πακέτο ggplot2 και data.table και αναλύονται οι δυνατότητες του εκάστοτε πακέτου που χρησιμοποιήθηκε.

Abstract

This thesis focuses on the study and analysis of labor force data derived from the related survey conducted by the Hellenic Statistical Authority, which is presented in detail in the first chapter. The main analysis is based on data collected in 2022, but additional data from 2018 to 2022 are used.

Initially, Descriptive Statistics are used to evaluate the employment situation in Greece. The study examines categories relevant to the whole labor force, such as age, education, region, nationality, and degree of urbanization. Subsequently, the analysis is divided into employment status: those who are employed and those who are unemployed, using different categories for each group. Finally, a comparative analysis is conducted on unemployment categories from 2018 to 2022 to understand how the phenomenon of unemployment changes over time. For the descriptive statistics, the R packages ggplot2 and data.table were used.

In the latter part of the thesis, the Logistic Regression model is fitted and presented for the data, where the dependence of unemployment on factors such as age, gender, education, and nationality is analyzed (first presenting the theoretical background of Logistic Regression). The last step involves using Classification and Cross-Validation to evaluate the model's capacity to adjust to new data.

In the final chapter, the code for the diagrams and tables created using the ggplot2 and data.table package is presented, and the functionalities of the packages that were used are analyzed.

Ευχαριστίες

Θα ήθελα να ευχαριστήσω την οικογένειά μου και τους φίλους μου για την αμέριστη υποστήριξή τους καθ' όλη τη διάρκεια εκπόνησης της εργασίας. Η πίστη σας, η αγάπη σας και η ενθάρρυνσή σας υπήρξαν σημαντικοί σύντροφοι όλα τα χρόνια των σπουδών μου.

Επιπλέον, θα ήθελα να ευχαριστήσω τον επιβλέποντα καθηγητή, Δημήτριο Φουσκάκη, για την ευκαιρία που μου έδωσε να συνεργαστώ μαζί του και τη συνεχή του καθοδήγηση.

1

[1]© (2024) Εθνικό Μετσόβιο Πολυτεχνείο. All rights Reserved. Απαγορεύεται η αντιγραφή, αποθήκευση και διανομή της παρούσας εργασίας, εξ ολοκλήρου ή τμήματος αυτής, για εμπορικό σκοπό. Επιτρέπεται η ανατύπωση, αποθήκευση και διανομή για σκοπό μη κερδοσκοπικό, εκπαιδευτικής ή ερευνητικής φύσης, υπό την προϋπόθεση να αναφέρεται η πηγή προέλευσης και να διατηρείται το παρόν μήνυμα. Ερωτήματα που αφορούν τη χρήση της εργασίας για κερδοσκοπικό σκοπό πρέπει να απευθύνονται προς το συγγραφέα. Οι απόψεις και τα συμπεράσματα που περιέχονται σ'αυτό το έγγραφο εκφράζουν το συγγραφέα και δεν πρέπει να ερμηνευτεί ότι αντιπροσωπεύουν τις επίσημες θέσεις του Εθνικού Μετσόβιου Πολυτεχνείου.

Στατιστική Ανάλυση Δεδομένων Εργατικού Δυναμικού με χρήση της R

Μαρία Ντάβου

30 Σεπτεμβρίου 2024

Περιεχόμενα

Κεφάλαιο 1

Εισαγωγή

1.1 Αντικείμενο και Σκοπός εργασίας

Η συγκεκριμένη εργασία έχει ως σκοπό την ανάλυση δεδομένων που σχετίζονται με το εργατικό δυναμικό της Ελλάδας, για το έτος 2022.

Τα δεδομένα προέρχονται από την αντίστοιχη έρευνα της Ελληνικής Στατιστικής Αρχής (ΕΛ.ΣΤΑΤ.), Έρευνα Εργατικού Δυναμικού (Ε.Ε.Δ). Η Έρευνα Εργατικού Δυναμικού αποτελεί την κύρια πηγή στοιχείων που σχετίζονται με την απασχόληση και την ανεργία. Η Ελληνική Στατιστική Αρχή αποτελεί την επίσημη εθνική στατιστική υπηρεσία, η οποία έχει ως σκοπό την συστηματική παραγωγή στατιστικών.

Η στατιστική ανάλυση δεδομένων που αφορούν το δυναμικό εργατικό βοηθά στην κατανόηση των παραγόντων που την επηρεάζουν, καθώς και στην ανάπτυξη στρατηγικών που θα αντιμετωπίζουν το φαινόμενο της ανεργίας και θα ενισχύουν τις ευκαρίες απασχόλησης.

1.2 Η Έρευνα Εργατικού Δυναμικού

Η Έρευνα Εργατικού Δυναμικού αποτελεί μια πανελλαδική έρευνα σε νοικοκυριά, η οποία διεξάγεται όλο το έτος και παράγει μηνιαία και τριμηνιαία αποτελέσματα. Συλλέγει στοιχεία σχετικά με την απασχόληση, την ανεργία και την εκπαίδευση. Βασικό της αποτέλεσμα είναι η κατάσταση απασχόλησης των ατόμων, που βρίσκονται σε εργάσιμη ηλικία (15 ετών και άνω) και διακρίνονται σε απασχολούμενοι, άνεργοι, μη ενεργοί. Η συγκεκριμένη έρευνα διεξάγεται σε όλες τις χώρες της Ευρώπης σύμφωνα με τους κανονισμούς της Eurostat.

Κατα τη διάρκεια των χρόνων, η συγκεκριμένη έρευνα έχει αλλάξει μορ-

φές. Από το 1974 έως το 1980 κάλυπτε μόνο αστικές και ημιαστικές περιοχές της Ελλάδας και ονομαζόταν Έρευνα Απασχόλησης. Το 1981 κάλυψε για πρώτη φορά ολόκληρη την χώρα, ακολουθώντας τις προϋποθέσεις των ερευνών που ακολουθούν όλες οι χώρες της Ευρωπαϊκής Οικονομικής Κοινότητας (ΕΟΚ). Μέχρι και το 1997 διεξαγόταν μια φορά το χρόνο, το δεύτερο τρίμηνο κάθε έτους. Από το 1998 διεξάγεται όλο το έτος και παράγει στοιχεία κάθε τρίμηνο του έτους. Λόγω του νέου Κανονισμού Πλαισίου Κοινωνικών Στατιστικών, 2019/1700, ο οποίος υιοθετήθηκε το 2021 και οδήγησε σε μεταβολές της έρευνας που αφορούσαν το σχεδιασμό της έρευνας, τον τρόπο συλλογής στοιχείων, τους ορισμούς και τις συλλεγόμενες πληροφορίες.

Η έρευνα αποτυπώνει την κατάσταση απασχόλησης των ερευνόμενων σε μια συγκεκριμένη περίοδο, η οποία αποτελεί την εβδομάδα αναφοράς. Οι εβδομάδες αναφοράς για κάθε τρίμηνο (και έτος) καθορίζονται από την Eurostat. Η εβδομάδα αναφοράς διαρκεί από Δευτέρα έως Κυριακή. Ο μήνας στον οποίο αντιστοιχούν, καθορίζεται από τον "κανόνα της Πέμπτης", δηλαδή η εβδομάδα αναφοράς αντιστοιχεί στον μήνα στον οποίο ανήκει η ημέρα Πέμπτη της εβδομάδας.

Μέσα από την Έρευνα Εργατικού Δυναμικού προκύπτουν στοιχεία που σχετίζονται με τα χαρακτηριστικά της κύριας εργασίας (ώρες, κλάδος, επάγγελμα κ.λπ.) και της αναζήτησης εργασίας (χρόνος αναζήτησης, εγγραφή σε γραφείο εύρεσης εργασίας κ.λπ.). Τέλος, αποτελεί πηγή για την συμμετοχή των ατόμων σε εκπαιδευτικές δραστηριότητες.

Η συγκεκριμένη έρευνα εξετάζεται μέσα από ερωτηματολόγιο, το οποίο χωρίζεται σε 8 θεματικές κατηγορίων ερωτήσεων. Τα τμήματα είναι τα εξής:

1. Χώρα γέννησης, υπηκοότητα, χρόνια παραμονής στην Ελλάδα και στοιχεία που αφορούν τους γονείς του ερωτόμενου.
2. Απασχόληση την εβδομάδα αναφοράς.
3. Χαρακτηριστικά της κύριας (ή μοναδικής) εργασίας.
4. Ύπαρξη και χαρακτηριστικά της τελευταίας εργασίας.
5. Αναζήτηση εργασίας.
6. Κύρια κατάσταση του (προσωπική θεώρηση της απασχόλησης).
7. Εκπαίδευση.

8. Πηγές συντήρησης του ατόμου και εισόδημα.

Το ερωτηματολόγιο συμπληρώνεται από τους ερευνητές της ΕΛ.ΣΤΑΤ. μέσω εξ΄ αποστάσεως ή δια ζώσης συνεντεύξεων των ερευνώμενων. Η μέθοδος συμπλήρωσης είναι με έντυπη φόρμα συμπλήρωσης (PAPI) ή μέσω ηλεκτρινικού μέσου, τάμπλετ/υπολογιστές (CAPI).

Κάθε νοικοκυριό που επιλέγεται στο δείγμα της έρευνας, ερευνάται για έξι διαδοχικά τρίμηνα. Κάθε τρίμηνο, το δείγμα της έρευνας εργατικού δυναμικού ανανεώνεται κατά το 1/6 ενώ τα υπόλοιπα 5/6 του δείγματος αποτελούνται από νοικοκυριά που έχουν ερευνηθεί σε προηγούμενο τρίμηνο. Η συνέντευξη με τα μέλη των νέων νοικοκυριών γίνεται με προσωπική συνέντευξη στην κατοικία του νοικοκυριού. Οι επόμενες συνεντεύξεις με ίδιο νοικοκυριό γίνονται είτε με συνέντευξη πρόσωπο με πρόσωπο είτε με τηλέφωνο. Επομένως, οι μονάδες χωρίζονται σε νέες, μονάδες στις οποίες το δείγμα ερευνάται για πρώτη φορά (1ο κύμα) και σε επαναλαμβανόμενες, μονάδες στις οποίες το δείγμα ερευνάται για δεύτερη έως έκτη φορά (κύματα). Ως νοικοκυριό ορίζονται:

- Τα άτομα που μένουν μόνα τους.
- Μια οικογένεια, με ή χωρίς παιδιά, με παππού-γιαγιά και τυχόν εσωτερική υπηρεσία.
- Μια οικογένεια στην οποία διανέμουν το πολύ 5 οικότροφοι.
- Δύο ή περισσότερα άτομα που έχουν νοικιάσει μια κατοικία (μαθητές, σπουδαστές, εργάτες, κ.λπ.). Αυτά τα άτομα αποτελούν ξεχωριστά νοικοκυριά.

Ως κατοικία ορίζεται ο χώρος που στεγάζει νοικοκυριά, δηλαδή μια οικία ή διαμέρισμα (ανεξάρτητα αν είναι κατοικημένο ή κενό), μια αποθήκη, καλύβα, παράγκα, σκηνή, βάρκα κ.λπ.., το οποίο κατοικείται κατά τη διεξαγωγή της Έρευνας. Εκτός Έρευνας τίθονται οι συλλογικές κατοικίες (ξενοδοχεία, στρατώνες, γηροκομεία) και επαγγελματικές στέγες (γραφεία, ιατρεία κ.λπ.).

Η επιλογή των νοικοκυριών που θα ερευνηθούν πραγματοποιείται με δισταδιακή στρωματοποιημένη δειγματοληψία. Αρχικά, η χώρα διαιρείται σε 206 στρώματα. Τα στρώματα δημιουργούνται κατανέμοντας τους Δήμους και τις Κοινότητες ανάλογα με τον συνολικό αριθμό των κατοίκων τους. Προκύπτουν τρείς ομάδες, η πρώτη περιλαμβάνει Συγκροτήματα και Δήμους με περισσότερους από 10.000 κατοίκους, Δήμοι και Κοινότητες με 2.000 έως 9.999 κατοίκους και Κοινότητες με έως

1.999 κατοίκους. Εξαίρεση απότελεί το Συγκρότημα της Αθήνας που διαιρείται σε 31 στρώματα και της Θεσσαλονίκης που διαιρείται σε 9 στρώματα. Στο πρώτο στάδιο της δειγματοληψίας όλες οι μονάδες επιφανείας έχουν επιλεγεί με πιθανότητα ανάλογη προς το μέγεθός τους κατά την τελευταία απογραφή που έχει πραγματοποιηθεί (στα δεδομένα του 2022 έχει χρησιμοποιηθεί η απογραφή του 2011) και με βάση τα ανανεωμένα δειγματοληπτικά πλαίσια (μεταξύ 2ου τριμήνου 2015 και 3ου τριμήνου 2016). Στο δεύτερο στάδιο της δειγματοληψίας επιλέχθηκε ένα συστηματικό δείγμα κατοικιών σε κάθε πρωτογενή δειγματοληπτική μονάδα. Σε όλα τα νοικοκυριά που διέμεναν στις επιλεγμένες κατοικίες ζητήθηκε να μετέχουν στην έρευνα. Στοιχεία συλλέγονται για όλα τα άτομα που αποτελούν μέλη των επιλεγμένων νοικοκυριών. Το συνολικό μέγεθος του τελικού δείγματος ανέρχεται σε 22.000 κατα μέσο όρο σε κάθε τρίμηνο.

Τα ερωτήματα της έρευνας μεταβάλλονται κάθε τρίμηνο. Υπάρχουν:

- Τριμηνιαία ερωτήματα, που γίνονται σε όλα τα κύματα της έρευνας.
- Ετήσια, συλλέγονται μια φορά το έτος.
- Διετή, συλλέγονται κάθε 2 χρόνια.
- Ad Hoc, συλλέγονται κάθε 8 χρόνια ανάλογα με το ειδικό θέμα.

Τα ετήσια, διετή και ad hoc ερωτήματα απαντώνται μόνο στο 1ο κύμα. Τα νοικοκυριά που ανήκουν σε επαναλαμβανόμενες μονάδες ερωτώνται μόνο για τις τριμηνιαίες ερωτήσεις, έστω και αν ερευνώνται για πρώτη φορά. Τα άτομα που ερευνώνται για πρώτη φορά απαντάνε σε όλα τα ερωτήματα, δημογραφικά, σχέσεις νοικοκυριού και τριμηνιαία. Αν ένα νοικοκυριό ερευνάται για δεύτερη φορά και παραπάνω, τότε σε όλα τα άτομα, ανεξαρτήτως ηλικίας, γίνεται έλεγχος για την παραμονή τους στο νοικοκυριό και τυχόν αλλαγή σχέσεων. Τα άτομα από 0-14 έτη και άνω των 75, δεν ερωτώνται τίποτα άλλο. Τα άτομα από 15-69 ετών ερωτώνται για όλα τα τριμηνιαία ερωτήματα. Τα άτομα από 70-74 ετών ερευνώνται για τα τριμηνιαία ερωτήματα αν στο προηγούμενο τρίμηνο άνηκαν στο εργατικό δυναμικό.

Σύμφωνα με το Διεθνή Οργανισμό Εργασίας (ILO) ως απασχολούμενοι ορίζονται τα άτομα που την εβδομάδα συμπλήρωσης του ερωτηματολογίου εργάστηκαν τουλάχιστον 1 ώρα ένταντι αμοιβής ή κέρδους ή τα άτομα με θέση εργασίας που δεν εργάζονται προσωρινά (π.χ. άδεια/εποχιακοί εργαζόμενοι/αργία). Άνεργοι θεωρούνται τα άτομα ηλικίας 15-74

που ήταν χωρίς εργασία την εβδομάδα αναφοράς (δηλαδή δεν θεωρούνται απασχολούμενοι), ήταν άμεσα διαθέσιμοι για εργασία ή αναζητούσαν ενεργά εργασία τις τελευταίες 4 εβδομάδες ή είχαν βρει εργασία την οποία θα αναλάμβαναν μέσα στους επόμενους μήνες.

Μη ενεργά χαρακτηρίζονται τα άτομα που δεν είναι ούτε απασχολούμενοι, ούτε άνεργοι.

1.3 Μεταβλητές που χρησιμοποιούνται

Οι μεταβλητές του συνόλου δεδομένων προέρχονται απο τις αντίστοιχες ερωτήσεις του ερωτηματολογίου που χρησιμοποιείται. Αποτελείται απο 119 μεταβλητές. Στο πλαίσιο της εργασίας θα χρησιμοποιηθούν συγκεκριμένες μεταβλητές, στις οποίες θα έχει περισσότερο ενδιαφέρον να εξεταστεί η κατάσταση απασχόλησης του πληθυσμού.

Το μέγεθος που μας ενδιαφέρει κυρίως είναι το πλήθος των ανέργων, το οποίο θα εξεταστεί σε διάφορες κατηγορίες. Σε κάποιες από αυτές τις κατηγορίες εξετάζεται το πλήθος των εργαζομένων και των ατόμων εκτός εργατικού δυναμικού.

Οι κατηγορίες είναι οι εξής:

- Η **κατάσταση απασχόλησης** του πληθυσμού, αποτελεί κατηγορική μεταβλητή με 3 κατηγορίες:
  - 1 = Εργαζόμενος
  - 2 = Άνεργος
  - 3 = Εκτός εργατικού δυναμικού
- Το **φύλο**, αποτελεί μια δίτιμη κατηγορική μεταβλητή:
  - 1 = Άνδρας
  - 2 = Γυναίκα
- Η **ηλικία**, αποτελεί μια συνεχή μεταβλητή.
- Το **ανώτατο επίπεδο εκπαίδευσης**, αποτελεί κατηγορική μεταβλητή με επτά κατηγορίες:
  - 1 = Χωρίς στοιχειώδη εκπαίδευση
  - 2 = Στοιχειώδη εκπαίδευση
  - 3 = Γυμνάσιο

  - 4 = Λύκειο
  - 5 = ΙΕΚ
  - 6 = Προπτυχιακό
  - 7 = Μεταπτυχιακό
- Η **περιφέρεια**, αποτελεί μια κατηγορική μεταβλητή με τους 13 δευτεροβάθμιους οργανισμούς τοπικής αυτοδιοίκησης που διαιρείται η Ελλάδα:
  - 1 = Ανατολική Μακεδονία και Θράκη
  - 2 = Κεντρική Μακεδονία
  - 3 = Δυτική Μακεδονία
  - 4 = Ήπειρος
  - 5 = Θεσσαλία
  - 6 = Ιόνια Νησιά
  - 7 = Δυτική Ελλάδα
  - 8 = Στερεά Ελλάδα
  - 9 = Αττική
  - 10 = Πελοπόννησος
  - 11 = Βόρειο Αιγαίο
  - 12 = Νότιο Αιγαίο
  - 13 = Κρήτη
- Η **υπηκόοτητα**, αποτελεί μια κατηγορική μεταβλητή με τρείς κατηγορίες:
  - 1 = Ελλάδα
  - 2 = Ευρωπαϊκή Ένωση
  - 3 = Άλλη Χώρα
- Η **αστικότητα**, αποτελεί μια κατηγορική μεταβλητή με πέντε κατηγορίες:
  - 1 = Περιφέρεια Πρωτευούσης
  - 2 = Πολεοδομικό Συγκρότημα (ΠΣ) Θεσσαλονίκης

- 3 = Αστική
- 4 = Ημιαστική
- 5 = Αγρότική

Οι παραπάνω μεταβλητές χρησιμοποιούνται τόσο στο σύνολο των ανέργων, των εργαζομένων, όσο και των ατόμων εκτός εργατικού δυναμικού. Οι επόμενες μεταβλητές εξετάζονται μόνο στο σύνολο των εργαζομένων και είναι οι εξής:

- Αν η κατάσταση απασχόλησης είναι **πλήρης ή μερική**, αποτελεί μια κατηγορική μεταβλητή με 2 κατηγορίες:
  - 1 = Πλήρης
  - 2 = Μερική
- Η **θέση εργασίας στην επιχείρηση** που εργάζονται, αποτελεί μια κατηγορική μεταβλητή με 3 κατηγορίες:
  - 1 = Αυτοαπασχολούμενοι
  - 2 = Μισθωτοί
  - 3 = Βοηθοί στην οικογενειακή επιχειρήση
- Το **Είδος της σύμβασης**, αποτελεί μια κατηγορική μεταβλητή με δύο κατηγορίες:
  - 1 = Είναι μόνιμη, με σύμβαση δημοσίου δικαίου
  - 2 = Είναι με σύμβαση αορίστου χρόνου
  - 2 = Είναι προσωρινή ή με σύμβαση ορισμένου χρόνο
- Οι **συνολικές εβδομαδιαίες ώρες εργασίας**, αποτελεί μια συνεχή μεταβλητή.
- Ο **κλάδος οικονομικής δραστηριότητας**, αποτελεί μια κατηγορική μεταβλητή με 21 κατηγορίες:
  - 01A = Γεωργία, Δασοκομία και Αλιεία
  - 02B = Ορυχεία και Λατομεία
  - 03C = Μεταποίηση
  - 04D = Παροχή ηλεκτρικού ρεύμαος, φυσικού αερίου, ατμού και κλιμαισμού

- 05E = Παροχή νερού, επεξεργασία λυμάτων, διαχείριση αποβλήτων και δραστηριότητες εξυγίανσης
- 06F = Κατασκευές
- 07G = Χονδρικό και λιανικό εμπόριο, επισκευή μηχανοκίνητων οχημάτων και μοτοσυκλετών
- 08H = Μεταφορά και αποθήκη
- 09J = Δραστηριότητες παροχής καταλυμάτων και υπηρεσιών εστίασης
- 10K = Ενημέρωση και επικοινωνία
- 11L = Χρηματοπιστωτικές και ασφαλιστικές υπηρεσίες
- 12M = Διαχείρισης ακίνητης περιουσίας
- 13N = Επαγγελματίες, επιστημονικές και τεχνικές δραστηριότητες
- 14O = Διοικητικές και υποστηρικτικές δραστηριότητες
- 15P = Δημόσια διοίκηση και άμυνα, υποχρεωτική κοινωνική ασφάλιση
- 16Q = Εκπαίδευση
- 17R = Δραστηριότητες σχετικές με την ανθρώπινη υγεία και κοινωνική μέριμνα
- 18S = Τέχνες, διασκέδαση και ψυχαγωγία
- 19T = Άλλες δραστηριότητες παροχής υπηρεσιών
- 20Y = Δραστηριότητες νοικοκυριών ως εργοδοτών, μη διαφοροποιημένες δραστηριότητες νοικοκυριών που αφορούν την παραγωγή αγαθών -και υπηρεσιών- για ίδια χρήση
- 21Z = Δραστηρίοτητες ετερόδικων οργανισμών και φορέων

- Το **επάγγελμα της κύριας απασχόλησης**, αποτελεί μια κατηγορική μεταβλητή με 7 κατηγορίες:
  - 1 = Ανώτερα διευθυντικά και διοικητικά στελέχη
  - 2 = Επαγγελματίες
  - 3 = Τεχνικοί και ασκούντες συναφή επαγγέλματα

  - 4 = Υπάλληλοι γραφείου
  - 5 = Απασχολούμενοι στην παροχή υπηρεσιών και πωλητές
  - 6 = Ειδικευμένοι γεωργοί, κτηνοτρόφοι, δασοκόμοι και αλιείς
  - 7 = Ειδικευμένοι τεχνίτες και ασκούντες συναφή επαγγέλματα
  - 8 = Χειριστές βιομηχανικών εγκαταστάσεων, μηχανημάτων και εξοπλισμού και συναρμολογητές (μονταδόροι)
  - 9 = Ανειδίκευτοι εργάτες, χειρωνάκτες και μικροεπαγγελματίες
  - Ο = Πρόσωπα μη δυνάμενα να καταταγούν

Οι μεταβλητές που ακολουθούν αφορούν αποκλειστικά το σύνολο των ανέργων.

- Ο **Λόγος διακοπής της τελευταίας εργασίας**, αποτελεί μια κατηγορική μεταβλητή με 7 κατηγορίες:
  - Ο = Διότι απολύθηκε ή έκλεισε η επιχείρηση για οικονομικούς λόγους
  - 1 = Διότι η εργασία ήταν περιορισμένης διάρκειας και τελείωσε
  - 2 = Φροντίζει μικρά παιδιά ή εξαρτώμενους ενήλικες
  - 3 = Για άλλους οικογενειακούς λόγους
  - 4 = Λόγω εκπαίδευσης ή επιμόρφωσης
  - 5 = Λόγω ασθένειας ή ανικανότητας
  - 6 = Λόγω συνταξιοδότησης
  - 7 = Άλλοι προσωπικοί λόγοι
  - 8 = Διότι στρατεύτηκε
  - 9 = Για άλλους λόγους
- Το **έτος της τελευταίας εργασίας**, αποτελεί μια συνεχή μεταβλητή.
- Η **διάρκεια αναζήτησης εργασίας**, αποτελεί μια κατηγορική μεταβλητή με 3 κατηγορίες:
  - 1 = Λιγότερο από 6 μήνες
  - 2 = 6-11 μήνες
  - 3 = 12 μηνες και άνω

- Η **Λήψη Επιδόματος**, αποτελεί μια κατηγορική μεταβλητή με δύο κατηγορίες:
  - 1 = Ναι
  - 2 = Όχι
- Η **Εγγραφή ως αναζητούσα/ων εργασίας**, αποτελεί μια κατηγορική μεταβλητή με δύο κατηγορίες:
  - 1 = Ναι
  - 2 = Όχι

Κεφάλαιο 2

Περιγραφική Στατιστική

Το πρώτο στάδιο σε μια στατιστική ανάλυση αποτελεί ο έλεγχος των δεδομένων. Μέσω του ελέγχου μπορούν να παρατηρηθούν έκτροπες, ελλιπείς και εσφαλμένες τιμές. Γνωρίζοντας ότι στο σύνολο των δεδομένων παρατηρείται κάποια από τις παραπάνω περιπτώσεις είμαστε σε θέση να προβούμε στον κατάλληλο χειρισμό τους και στην αντιμετώπιση τους. Αυτό επιτυγχάνεται μέσω της Περιγραφικής Στατιστικής. Στον παρόν κεφάλαιο, πραγματοποιείται η Περιγραφική Στατιστική των δεδομένων με την βοήθεια της Στατιστικής Γλώσσας R και παρουσιάζονται τα αποτελέσματα μέσω πινάκων και διαγραμμάτων. Οι γραφικές μέθοδοι της Περιγραφικής Στατιστικής δημιουργούνται με το πακέτο **ggplot2** της R, ενώ τα σύνολα των διάφορων κατηγοριών έχουν υπολογιστεί με το πακέτο **data.table**.

Η ανάλυση χωρίζεται στις 3 κατηγορίες ατόμων. Αρχικά, εξετάζεται η συνολική κατάσταση απασχόλησης, η οποία αποτελείται από τους εργαζόμενους, τους ανέργους και τα άτομα εκτός εργατικού δυναμικού, σε μεταβλητές που είναι κοινές και για τις 3 κατηγορίες. Στην συνέχεια εξετάζονται οι εργαζόμενοι σε μεταβλητές που αφορούν αποκλειστικά αυτή την κατηγορία, και τέλος εξετάζονται οι άνεργοι με αντίστοιχο τρόπο.

Στο τέλος του κεφαλαίου, εξετάζεται η εξέλιξη της ανεργίας από τα έτη 2018 έως 2022.

2.1 Η Κατάσταση Απασχόλησης για το 2022

Το ποσοστό της ανεργίας και της απασχόλησης υπολογίζεται μεταξύ του συνολικού αριθμού εργαζομένων και ανέργων, σύμφωνα με τον ορισμό του Διεθνή Οργανισμού Εργασίας (ILO). Το ποσοστό των ατόμων εκ-

τός εργατικού δυναμικού υπολογίζεται χρησιμοποιώντας τον σύνολο των ατόμων που ερωτήθηκαν.

Το συνολικό ποσοστό ανεργίας για την Ελλάδα το 2022 ανέρχεται στο 11,7%, όπως παρατηρείται στον Πίνακα 2.1. Σύμφωνα με την Eurostat, Πίνακας 2.2, η συνολική ανεργία στις 27 χώρες της Ευρωπαϊκής Ένωσης είναι στο 6,2% και στην ευρύτερη περιοχή της Ευρώπης στο 6,8%. Η Ελλάδα βρίσκεται στην 2η θέση, ενώ στην 1η θέση η Ισπανία με 12,9%. Στην τελευταία θέση βρίσκεται η Τσεχία με 2,2%.

Κατάσταση	**Σύνολο**	**Ποσοστό**
Εργαζόμενοι	64085	88,3%
Άνεργοι	8510	11,7%
Εκτός Εργατικού Δυναμικού	89786	55,3%

Πίνακας 2.1: Σύνολο και Ποσοστό εργαζομένων, ανέργων και ατόμων εκτός εργατικού δυναμικού

Χώρα	**Ποσοστό**
Ισπανία	12,9%
Ελλάδα	11,7%
Ευρώπη	6,8%
Ευρωπαϊκή Ένωση	6,2%
Τσεχία	2,2%

Πίνακας 2.2: Ποσοστά ανεργίας στην Ευρώπη σύμφωνα με την Eurostat για το 2022

2.1.1 Ως προς το Φύλο

Εξετάζοντας την κατάσταση απασχόλησης ανάμεσα στους άνδρες και τις γυναίκες παρατηρείται, όπως φαίνεται στο Σχήμα 2.1, μεγαλύτερη απορρόφηση στους άνδρες, με το 91,1% να εργάζεται και μεγαλύτερη ανεργία στις γυναίκες, στο 15,1% σε αντίθεση με το 8,9% των ανδρών. Παρατηρούμε ότι τα άτομα εκτός εργατικού δυναμικού συγκροτούνται κυρίως από γυναίκες με ποσοστό 51,2% και εν συνεχεία οι άνδρες με 48,8% επί του συνολικού αριθμού των ερωτηθέντων, σε άνδρες και γυναίκες αντίστοιχα.

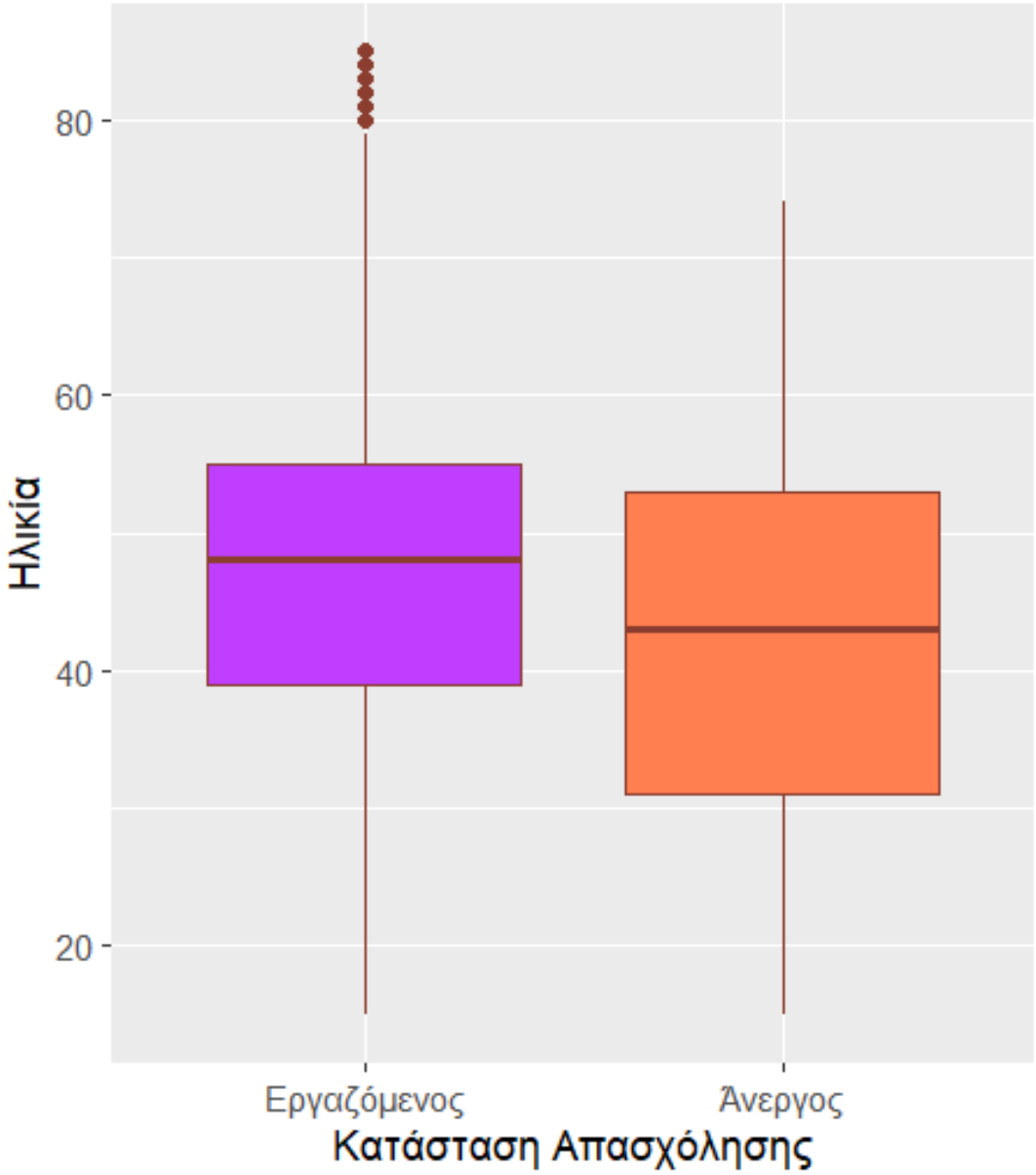
Σχήμα 2.1: Κατάσταση Απασχόλησης ανά Φύλο

2.1.2 Ως προς την Ηλικία

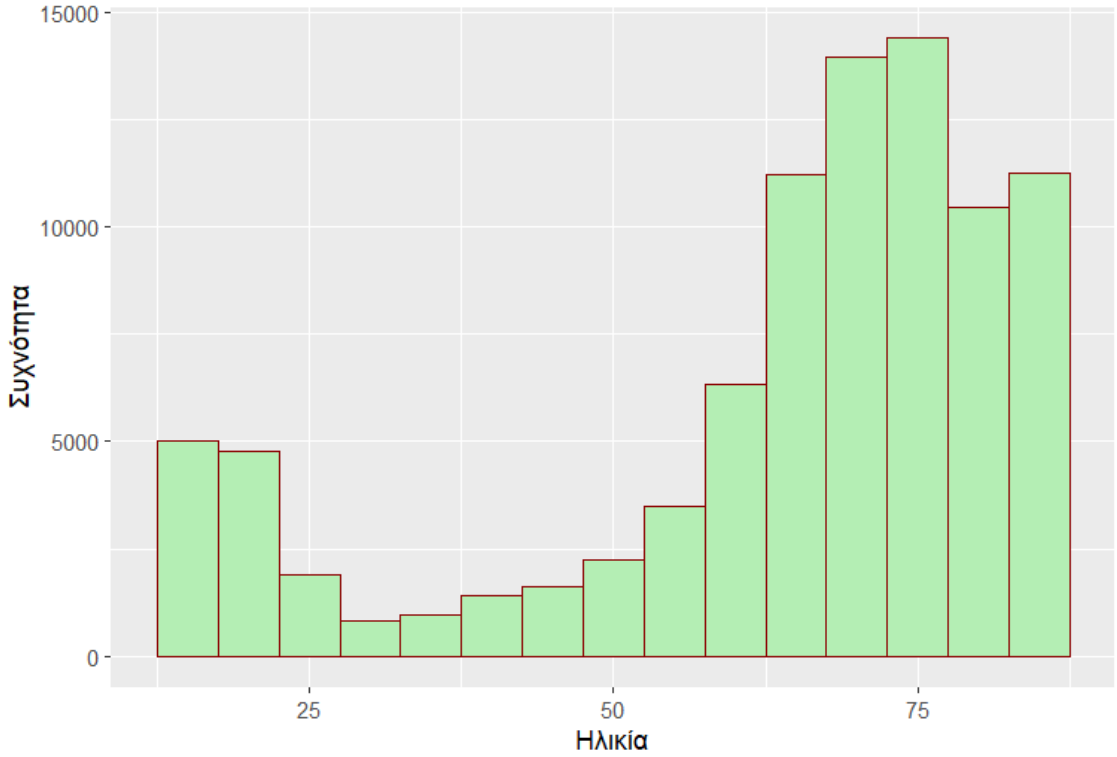
Η μέση ηλικία των ερωτηθέντων είναι τα 51,4 έτη. Όσον αφορά την μέση ηλικία των εργαζομένων είναι στα 46,64 έτη, των ανέργων στα 42,22 έτη και των ατόμων εκτός εργατικού δυναμικού στα 63,2 έτη, αυτό παρατηρείται στον Πίνακα 2.3, στο Σχήμα 2.2 και 2.3. Επομένως, τα άτομα εκτός εργατικού δυναμκού αποτελούνται κυρίως από γηραιότερα άτομα, σε σχέση με τα άτομα που είναι ενεργά όσον αφορά την εργασία (είτε εργάζονται, είτε είναι άνεργα αλλά διαθέσιμα για εργασία). Λόγω της υψηλής μέσης ηλικίας των ατόμων εκτός εργατικού δυναμικού και συνδυαστικά με το ιστόγραμμα 2.3 παρατηρούμε ότι το μεγαλύτερο μέρος αυτών των ατόμων έχει υψηλές ηλικίες. Επομένως, συμπεραίνουμε ότι τα περισσότερα από τα άτομα που ανήκουν στην κατηγορία εκτός εργατικού δυναμικού βρίσκονται σε ηλικία συνταξιοδότησης, η οποία είναι τα 62 ή 67 έτη στην Ελλάδα.

Κατάσταση Απασχόλησης	Μέση ηλικία
Εργαζόμενοι	46,64
Άνεργοι	42,22
Εκτός Εργατικού Δυναμικού	63,2

Πίνακας 2.3: Μέση ηλικία ανά Κατάσταση Απασχόλησης



Σχήμα 2.2: Θηκόγραμμα ηλικίας εργαζομένων και ανέργων



Σχήμα 2.3: Ιστόγραμμα ηλικίας ατόμων εκτός εργατικού δυναμικού

Εξετάζοντας την μέση ηλικία του κάθε φύλου ανά κατάσταση απασχόλησης προκύπτουν τα αποτελέσματα του Πίνακα 2.4. Παρατηρούμε πως δεν υπάρχει ιδιαίτερη ηλικιακή διαφορά ανάμεσα στα δύο φύλα, με τις γυναίκες να είναι ελαφρώς νεότερες (λιγότερο από ένα έτος) σε κάθε κατάσταση. Επιβεβαιώνεται ξανά ότι τα άτομα που αποτελούν την κατηγορία εκτός εργατικού δυναμικού συντελούν τον γηραιότερο πληθυσμό, ανεξαρτήτως του φύλου.

Φύλο	**Κατάσταση Απασχόλησης**	**Μέση ηλικία**
Άνδρας	Εργαζόμενος	46,81
Γυναίκα	Εργαζόμενη	46,41
Άνδρας	Άνεργος	42,35
Γυναίκα	Άνεργη	42,11
Άνδρας	Εκτός Εργατικού Δυναμικού	63,75
Γυναίκα	Εκτός Εργατικού Δυναμικού	62,69

Πίνακας 2.4: Μέση ηλικία ανά Φύλο και ανά Κατάσταση Απασχόλησης

2.1.3 Ως προς το Ανώτατο Επίπεδο Εκπαίδευσης

Αν εξετάσουμε την κατάσταση απασχόλησης στην κατηγορία που αφορά το ανώτατο επίπεδο εκπαίδευσης παρατηρούμε από τους Πίνακες 2.5,

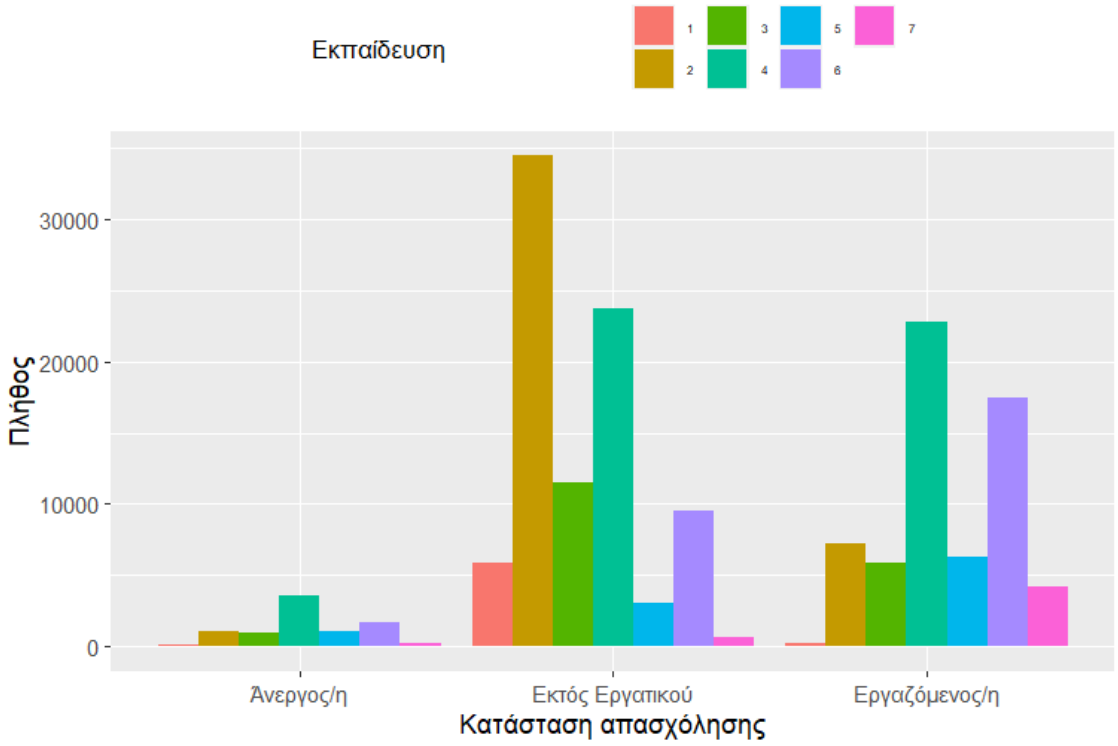
2.6 και Σχήμα 2.4 πως η μεγαλύτερη απορρόφηση υπάρχει στα άτομα που έχουν ολοκληρώσει το Λύκειο, το 35,6% και ακολουθούν τα άτομα με προπτυχιακές σπουδές, με 27,2% και στην τελευταία θέση τα άτομα που δεν έχουν λάβει καμία στοιχειώδη εκπαίδευση, με 0,4%. Όσον αφορά την ανεργία, μεγαλύτερο ποσοστό παρατηρείται στους απόφοιτους Λυκείου, με 41,3% και στα άτομα με προπτυχιακές σπουδές, με 19,1%. Τέλος, το μεγαλύτερο ποσοστό των ατόμων εκτός εργατικού δυναμικού έχει λάβει μόνο την στοιχειώδη εκπαίδευση, συγκεκριμένα το 38,8% , ενώ το μικρότερο ποσοστό αποτελούν τα άτομα με μεταπτυχιακές σπουδές, με 0,7%. Παρατηρούμε ότι το ποσοστό των ατόμων μεγαλύτερης ηλικίας, όπως είναι τα άτομα εκτός εργατικού δυναμικού, είναι μικρότερο όσο ανεβαίνουμε τις βαθμίδες της εκπαίδευσης.

Επίπεδο Εκπαίδευσης	**Εργαζόμενοι**		**Άνεργοι**	
	Σύνολο	Ποσοστό	Σύνολο	Ποσοστό
1 = Χωρίς στοιχειώδη εκπαίδευση	247	0,4%	73	0,85%
2 = Στοιχειώδη εκπαίδευση	7216	11,2%	1025	12,0%
3 = Γυμνάσιο	5880	9,2%	1025	12,0%
4 = Λύκειο	22805	35,6%	3540	41,3%
5 = ΙΕΚ	6270	9,8%	1032	12%
6 = Προπτυχιακό	17458	27,2%	1637	19,1%
7 = Μεταπτυχιακό	4209	6,6%	236	2,75%

Πίνακας 2.5: Σύνολο και Ποσοστό εργαζομένων και ανέργων ανά Ανώτατο Επίπεδο Εκπαίδευσης

Εκτός Εργατικού Δυναμικού	**Σύνολο**	**Ποσοστό**
1 = Χωρίς στοιχειώδη εκπαίδευση	5868	6,6%
2 = Στοιχειώδη εκπαίδευση	34489	38,8%
3 = Γυμνάσιο	11563	13,0%
4 = Λύκειο	23707	26,7%
5 = ΙΕΚ	3093	3,5%
6 = Προπτυχιακό	9532	10,7%
7 = Μεταπτυχιακό	655	0,7%

Πίνακας 2.6: Σύνολο και Ποσοστό ατόμων εκτός εργατικού δυναμικού ανά Ανώτατο Επίπεδο Εκπαίδευσης



Σχήμα 2.4: Κατάσταση απασχόλησης ανά Ανώτατο Επίπεδο Εκπαίδευσης

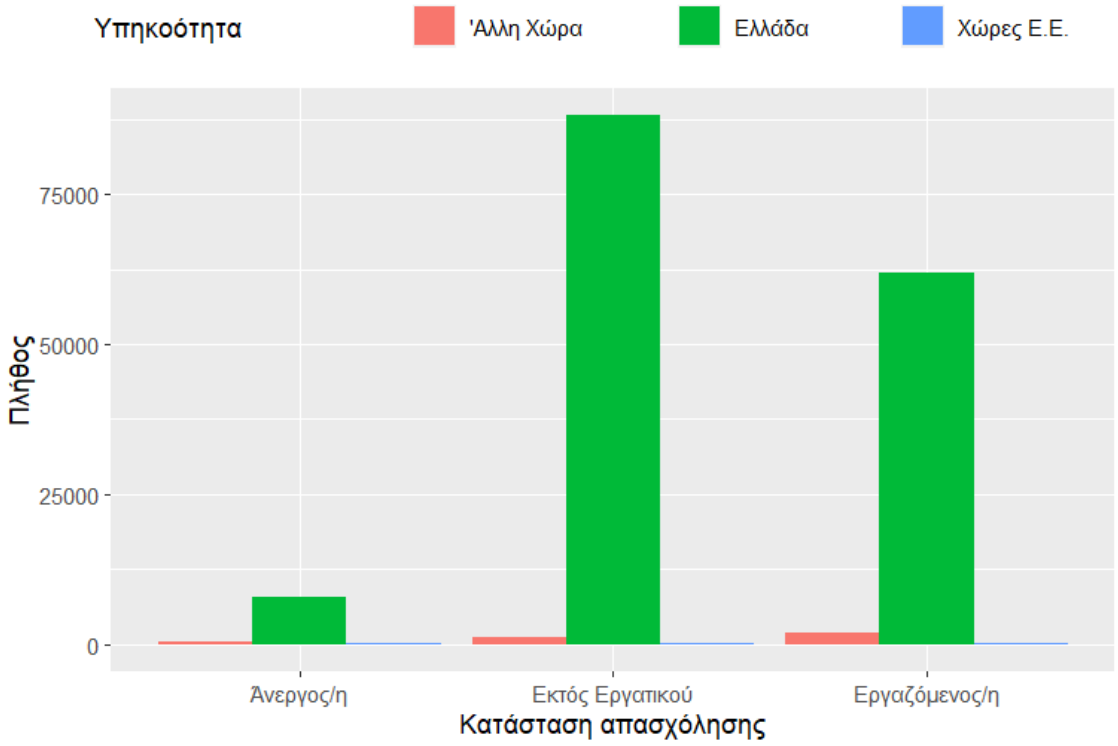
Σύμφωνα με τα αποτελέσματα στον Πίνακα 2.7, όπου εξετάζει την κατάσταση απασχόλησης ανά ανώτατο επίπεδο εκπαίδευσης και εστιάζει ανάμεσα στα δύο φύλα, παρατηρούμε ότι οι περισσότερες γυναίκες ανήκουν στα άτομα εκτός εργατικού δυναμικού που έχουν ολοκληρώσει την στοιχειώδη εκπαίδευση και το λύκειο και μεγάλο ποσοστό υπάρχει στους άνδρες εκτός εργατικού δυναμικού που έχουν ολοκληρώσει την στοιχειώδη εκπαίδευση. Ενώ οι περισσότεροι άνδρες ανήκουν σε εργαζόμενα άτομα και είναι απόφοιτοι λυκείου. Περισσότερες φαίνεται να είναι οι γυναίκες με προπτυχιακό τίτλο και στις 3 καταστάσεις απασχόλησης. Συνολικά το υψηλό ποσοστό των ατόμων εκτός εργατικού δυναμικού που έχουν ολοκληρώσει την στοιχειώδη εκπαίδευση είναι κάτι που αναμένουμε από την υψηλή ηλικία των ερωτηθέντων.

Εκπαίδευση	Κατάσταση Απασχόλησης	Άνδρας	Γυναίκα
Χωρίς στοιχειώδη εκπαίδευση	Εργαζόμενος	162	85
Χωρίς στοιχειώδη εκπαίδευση	Άνεργος	54	19
Χωρίς στοιχειώδη εκπαίδευση	Εκτός Εργατικού Δυναμικού	1947	3921
Στοιχειώδη εκπαίδευση	Εργαζόμενος	4152	3064
Στοιχειώδη εκπαίδευση	Άνεργος	513	512
Στοιχειώδη εκπαίδευση	Εκτός Εργατικού Δυναμικού	13777	20712
Γυμνάσιο	Εργαζόμενος	3983	1897
Γυμνάσιο	Άνεργος	452	515
Γυμνάσιο	Εκτός Εργατικού Δυναμικού	5490	6073
Λύκειο	Εργαζόμενος	14446	8359
Λύκειο	Άνεργος	1530	2010
Λύκειο	Εκτός Εργατικού Δυναμικού	10046	13661
ΙΕΚ	Εργαζόμενος	3240	3030
ΙΕΚ	Άνεργος	330	702
ΙΕΚ	Εκτός Εργατικού Δυναμικού	1330	1763
Προπτυχιακό	Εργαζόμενος	8335	9123
Προπτυχιακό	Άνεργος	599	1038
Προπτυχιακό	Εκτός Εργατικού Δυναμικού	4737	4795
Μεταπτυχιακό	Εργαζόμενος	2056	2153
Μεταπτυχιακό	Άνεργος	99	137
Μεταπτυχιακό	Εκτός Εργατικού Δυναμικού	357	298

Πίνακας 2.7: Κατάσταση Απασχόλησης και Φύλο ανά Επίπεδο Εκπαίδευσης

2.1.4 Ως προς την Υπηκοότητα

Εξετάζοντας την κάθε κατάσταση απασχόλησης με την υπηκοότητα στο Σχήμα 2.5, παρατηρούμε ότι με μεγάλη διαφορά η μεγαλύτερη απορρόφηση, ανεργία και άτομα εκτός εργατικού δυναμικού συντελούνται από άτομα με Ελληνική υπηκόοτητα. Ενώ ακολουθούν τα άτομα με υπηκοότητα από χώρα της Ευρωπαϊκής Ένωσης και τέλος τα άτομα με υπηκοότητα απο Άλλη Χώρα. Τα περισσότερα άτομα με Ελληνική υπηκοότητα ανήκουν στην κατάσταση εκτός εργατικού δυναμικού, τα περισσότερα άτομα με υπηκοότητα από χώρα της Ευρωπαϊκής Ένωσης ανήκουν στα εργαζόμενα άτομα και τέλος, τα άτομα με υπηκοότητα από Άλλη Χώρα ανήκουν στα εκτός εργατικού δυναμικου. Επομέμως, η μεγαλύτερη απορρόφηση παρατηρείται στα άτομα με Ελληνική υπηκοότητα, με μεγάλη διαφορά από τις υπόλοιπες δύο υπηκοότητες.



Σχήμα 2.5: Κατάσταση απασχόλησης ανά Υπηκοότητα

2.1.5 Ως προς το Βαθμό Αστικοποίησης

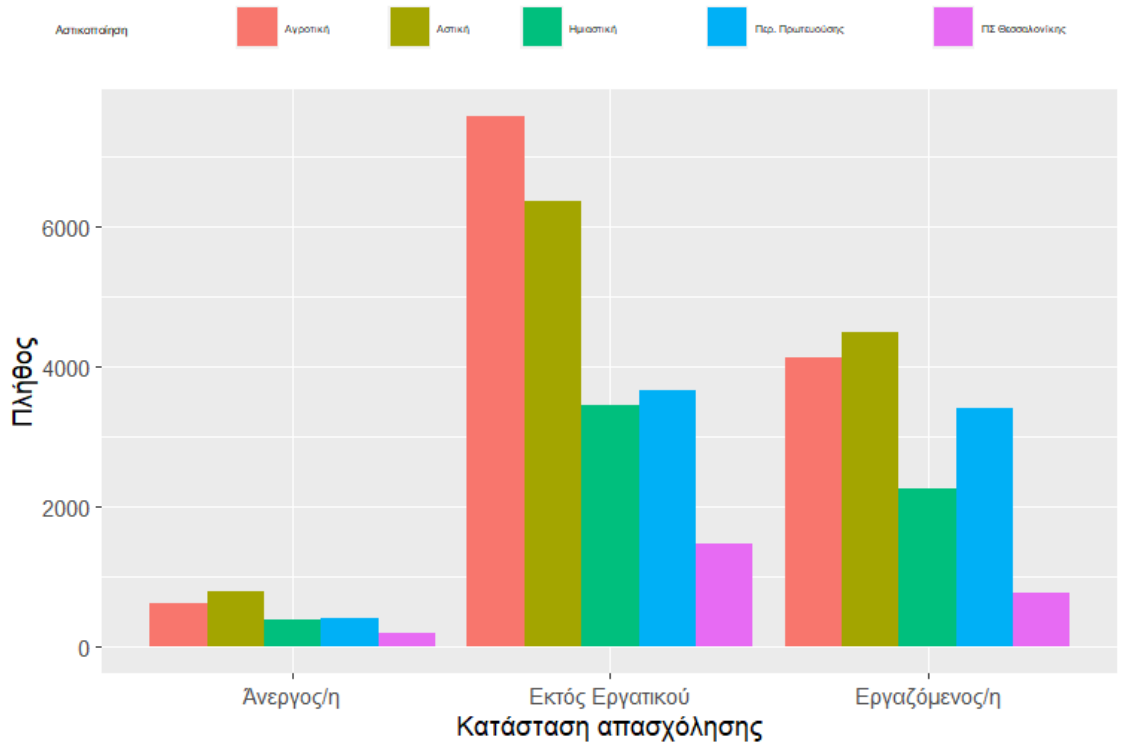
Από τον Πίνακα 2.8, 2.9 και το Σχήμα 2.6 τα μεγαλύτερα ποσοστά απορρόφησης και ανεργίας παρατηρούνται στις Αστικές περιοχές με ποσοστά 29,8% και 35,9% αντίστοιχα. Ενώ το 33,7% των ατόμων εκτός εργατικού δυναμικού βρίσκεται σε Αγρότική περιοχή. Τα μικρότερα ποσοστά εμφανίζονται στο Πολεοδομικό Συγκρότημα της Θεσσαλονίκης με ποσοστά 5,0%, 8,3% και 6,5% αντίστοιχα. Συμπεραίνουμε ότι τα άτομα μεγαλύτερων ηλικιακών ομάδων, τα άτομα εκτός εργατικού δυναμικού, πρότιμούν τις αγρότικες περιοχές σε σχέση με τα άτομα που θεωρούνται ενεργά (είτε εργάζονται, είτε όχι) τα οποία πρότιμούν τις αστικές περιοχές.

Αστικότηρα	**Εργαζόμενοι**		**Άνεργοι**	
	Σύνολο	Ποσοστό	Σύνολο	Ποσοστό
Αστική	4478	29,8%	777	33,1%
Αγρότική	4128	27,5%	605	25,8%
Περιφέρεια Πρωτευούσης	3407	22,7%	396	16,9%
Ημιαστική	2245	15%	386	16,5%
ΠΣ Θεσσαλονίκης	755	5%	180	7,7%

Πίνακας 2.8: Σύνολο και Ποσοστό εργαζομένων και ανέργων ανά Βαθμό Αστικοποίησης

Εκτός Εργατικού Δυναμικού	Σύνολο	Ποσοστό
Αγρότική	7570	33,7%
Αστική	6347	28,2%
Περιφέρεια Πρωτευούσης	3662	16,3%
Ημιαστική	3433	15,3%
ΠΣ Θεσσαλονίκης	1461	6,5%

Πίνακας 2.9: Σύνολο και Ποσοστό ατόμων εκτός εργατικού δυναμικού ανά Βαθμό Αστικοποίησης



Σχήμα 2.6: Κατάσταση απασχόλησης ανά Βαθμό Αστικοποίησης

Εξετάζοντας και την κατάσταση απασχόλησης ανάμεσα στα δύο φύλα και τον βαθμό αστικοποίησης συμπεραίνουμε, Πίνακας 2.10, ότι οι περισσότεροι εργαζόμενοι άνδρες βρίσκονται σε αγρότικές και αστικές περιοχές, ενώ οι γυναίκες σε αστικές και στην Περιφέρεια Πρωτευούσης. Σε αστικές και αγρότικές περιοχές βρίσκεται και το μεγαλύτερο ποσοστό ατόμων εκτός εργατικού δυναμικού. Το Πολεοδομικό Συγκρότημα Θεσσαλονίκης περιέχει τα μικρότερα ποσοστά με τους άνδρες να υπερτερούν ως εργαζόμενοι, ενώ οι γυναίκες παρουσιάζονται περισσότερες ως άνεργες και εκτός εργατικού δυναμικού.

Περιοχή	Κατάσταση Απασχόλησης	Άνδρας	Γυναίκα
Περιφέρεια Πρωτευούσης	Εργαζόμενος	1843	1564
Περιφέρεια Πρωτευούσης	Άνεργος	178	218
Περιφέρεια Πρωτευούσης	Εκτός Εργατικού Δυναμικού	1495	2167
ΠΣ Θεσσαλονίκης	Εργαζόμενος	413	342
ΠΣ Θεσσαλονίκης	Άνεργος	68	112
ΠΣ Θεσσαλονίκης	Εκτός Εργατικού Δυναμικού	591	870
Αστική	Εργαζόμενος	2478	2000
Αστική	Άνεργος	276	501
Αστική	Εκτός Εργατικού Δυναμικού	2685	3662
Ημιαστική	Εργαζόμενος	1300	945
Ημιαστική	Άνεργος	170	216
Ημιαστική	Εκτός Εργατικού Δυναμικού	1472	1961
Αγρότική	Εργαζόμενος	2519	1609
Αγρότική	Άνεργος	303	302
Αγρότική	Εκτός Εργατικού Δυναμικού	3327	4243

Πίνακας 2.10: Κατάσταση Απασχόλησης και Φύλο ανά Βαθμό Αστικοποίησης

Εξετάζοντας πως κυμαίνεται το ανώτατο επίπεδο εκπαίδευσης σε σχέση με τον βαθμό αστικοποίησης μιας περιοχής προκύπτουν τα αποτελέσματα του Πίνακα 2.11, 2.12 και 2.13. Παρατηρούμε ότι η Περιφέρεια Πρωτευούσης αποτελείται κυρίως από απόφοιτους Δευτεροβάθμιας (Λύκειο) και Τριτοβάθμιας εκπαίδευσης (Προπτυχιακό). Ομοίως και το Πολεοδομικό Συγκρότημα της Θεσσαλονίκης, το οποίο έχει και υψηλό ποσοστό στα άτομα που έχουν ολοκληρώσει την στοιχειώδη εκπαίδευση, ποσοστό το οποίο εμφανίζεται υψηλότερο από το αντίστοιχο για τα άτομα που έχουν ολοκληρώσει κάποιο προπτυχιακό πρόγραμμα. Στις Αστικές περιοχές το μεγαλύτερο ποσοστό των ατόμων προκύπτει ότι είναι απόφοτοιτοι Λυκείου, ακολουθούν τα άτομα με προπτυχιακό τίτλο και τα άτομα που έχουν ολοκληρώσει την στοιχειώδη εκπαίδευση. Η κατανομή των ατόμων στις ημιαστικές περιοχές φαίνεται να είναι παρόμοια, με άτομα που έχουν ολοκληρώσει το Λύκειο να προηγούνται και να ακολουθούν τα άτομα που έχουν ολοκληρώσει την στοιχειώδη εκπαίδευση. Τέλος, στις αγρότικές περιοχές τα περισσότερα άτομα έχουν ολοκληρώσει την στοιχειώδη εκπαίδευση και ακολουθούν οι απόφοιτοι Λυκείου.

Από όλα τα παραπάνω προκύπτει ότι οι περισσότεροι άνεργοι άνδρες και γυναίκες βρίσκονται σε αστικές και αγρότικές περιοχές και είναι απόφοιτοι δευτεροβάθμιας εκπαίδευσης.

Εκπαίδευση	Π.Πρωτευούσης		ΠΣ Θεσσαλονίκης	
	Σύνολο	Ποσοστό	Σύνολο	Ποσοστό
Χωρίς στοιχειώδη εκπαίδευση	85	1,2%	76	3,2%
Στοιχειώδη εκπαίδευση	858	11,7%	543	23,1%
Γυμνάσιο	574	7,8%	209	8,9%
Λύκειο	2442	33,3%	788	33,5%
ΙΕΚ	818	11,1%	143	6,1%
Προπτυχιακό	1962	26,7%	472	20,1%
Μεταπτυχιακό	598	8,2%	119	5,1%

Πίνακας 2.11: Σύνολο και Ποσοστό ατόμων ανά Ανώτατο Επίπεδο Εκπαίδευσης και Βαθμό Αστικοποίησης

Εκπαίδευση	Αστική		Ημιαστική	
	Σύνολο	Ποσοστό	Σύνολο	Ποσοστό
Χωρίς στοιχειώδη εκπαίδευση	277	2,4%	255	4,3%
Στοιχειώδη εκπαίδευση	2224	19,4%	1845	31%
Γυμνάσιο	1323	11,6%	765	12,8%
Λύκειο	3826	33,5%	1866	31,3%
ΙΚΕ	833	7,3%	301	5,1%
Προπτυχιακό	2612	22,8%	861	14,4%
Μεταπτυχιακό	346	3%	66	1,1%

Πίνακας 2.12: Σύνολο και Ποσοστό ατόμων ανά Ανώτατο Επίπεδο Εκπαίδευσης για Αστική και Ημιαστική περιοχή

Εκπαίδευση	Αγρότική	
	Σύνολο	Ποσοστό
Χωρίς στοιχειώδη εκπαίδευση	860	7,2%
Στοιχειώδη εκπαίδευση	4979	41,9%
Γυμνάσιο	1502	12,6%
Λύκειο	3019	25,4%
ΙΚΕ	491	4,1%
Προπτυχιακό	981	8,3%
Μεταπτυχιακό	56	0.5%

Πίνακας 2.13: Σύνολο και Ποσοστό ατόμων ανά Ανώτατο Επίπεδο Εκπαίδευσης για Αγρότική περιοχή

2.1.6 Ως προς την Περιφέρεια

Εξετάζουμε την κατάσταση απασχόλησης στις 13 Περιφερειακές ενότητες της Ελλάδας και στους Πίνακες 2.14, 2.15 και στο Σχήμα2.7 βλέπουμε τα αποτελέσματα που προκύπτουν. Τα ποσοστά που προκύπτουν αφορούν τους εργαζομένους, τους ανέργους και τα άτομα εκτός εργατικού

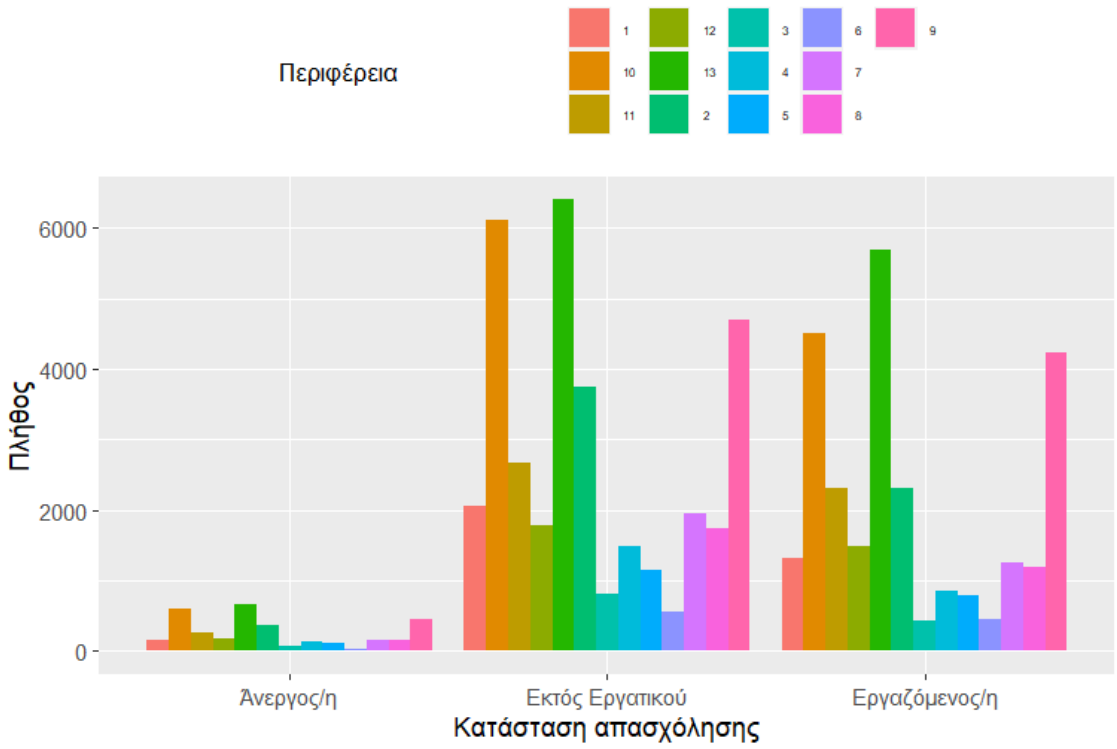
δυναμικού. Τα μεγαλύτερα ποσοστά απασχόλησης παρατηρούνται στην Κρήτη με 21,2%, στην Πελοπόννησο με 16,8% και στην Αττική με 15,7%, ενώ η μικρότερη παρατηρείται στα Νησιά του Ιονίου με 1,8%. Αντίστοιχα, η μεγαλύτερη ανεργία παρατηρείται στην Κρήτη με 19,5%, στην Πελοπόννησο με 18,2% και στην Αττική με 13,6%. Το μικρότερο ποσοστό ανεργίας παρατηρείται στην Περιφέρεια των Ιονίων Νήσων. Τέλος, τα περισσότερα άτομα εκτός εργατικού δυναμικού βρίσκονται στην Κρήτη με 18,2%, στην Πελοπόννησο με 17,3% και στην Αττική με 13,3%, ενώ το μικρότερο ποσοστό εμφανίζεται στα Ιόνια Νησιά. Συμπεραίνουμε ότι η περιφερειακή ενότητα της Κρήτης, της Πελοποννήσου και της Αττικής παρουσιάζουν τα μεγαλύτερα ποσοστά και στις τρείς καταστάσεις απασχόλησης.

Περιφέρεια	**Εργαζόμενοι**		**Άνεργοι**	
	Σύνολο	Ποσοστό	Σύνολο	Ποσοστό
1 = Ανατολική Μακεδονία και Θράκη	1311	4,9%	161	4,8%
2 = Κεντρική Μακεδονία	2313	8,6%	366	11,0%
3 = Δυτική Μακεδονία	423	1,6%	70	2,1%
4 = Ήπειρος	846	3,1%	130	3,9%
5 = Θεσσαλία	793	3,0%	122	3,6%
6 = Ιόνια Νησιά	461	1,7%	28	0,8%
7 = Δυτική Ελλάδα	1248	4,7%	158	4,7%
8 = Στερεά Ελλάδα	1187	4,4%	162	4,8%
9 = Αττική	4235	15,8%	457	13,6%
10 = Πελοπόννησος	4517	16,8%	610	18,2%
11 = Βόρειο Αιγαίο	2306	8,6%	263	7,9%
12 = Νότιο Αιγαίο	1490	5,6%	168	5,0%
13 = Κρήτη	5697	21,2%	655	19,6%

Πίνακας 2.14: Σύνολο και Ποσοστό εργαζομένων και ανέργων ανά Περιφέρεια

Εκτός Εργατικού Δυναμικού	Σύνολο	Ποσοστό
1 = Ανατολική Μακεδονία και Θράκη	2050	5,8%
2 = Κεντρική Μακεδονία	3739	10,6%
3 = Δυτική Μακεδονία	813	2,3%
4 = Ήπειρος	1494	4,2%
5 = Θεσσαλία	1141	3,2%
6 = Ιόνια Νησιά	556	1,6%
7 = Δυτική Ελλάδα	1961	5,6%
8 = Στερεά Ελλάδα	1747	5,0%
9 = Αττική	4695	13,4%
10 = Πελοπόννησος	6114	17,4%
11 = Βόρειο Αιγαίο	2666	7,6%
12 = Νότιο Αιγαίο	1788	5,1%
13 = Κρήτη	6413	18,2%

Πίνακας 2.15: Σύνολο και Ποσοστό ατόμων εκτός εργατικού δυναμικού ανά Περιφέρεια



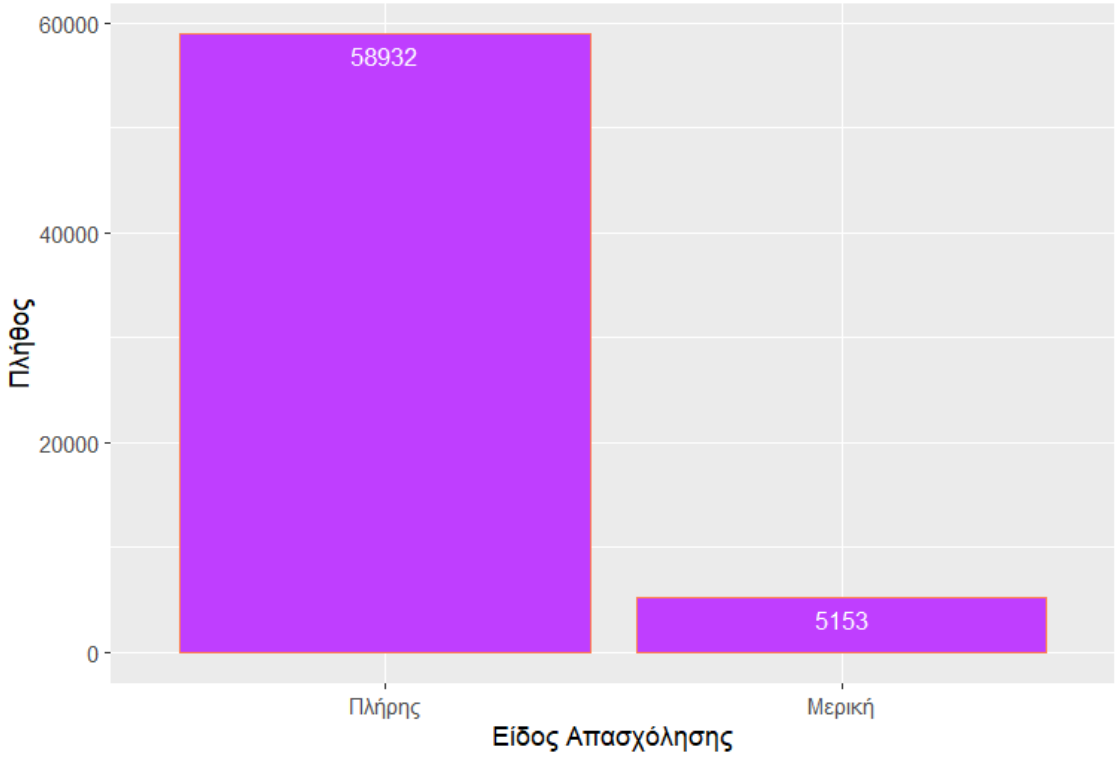
Σχήμα 2.7: Κατάσταση απασχόλησης ανά Περιφέρεια

2.2 Σχετικά με τους Εργαζομένους

Στη παρούσα παράγραφο εξετάζεται το σύνολο των Εργαζομένων ως προς τον οικονομικό κλάδο εργασίας τους, την κύρια δραστηριότητα, την κύρια θέση, το είδος της απασχόλησης, τις ώρες εργασίας και το είδος της σύμβασης τους.

2.2.1 Είδος Απασχόλησης

Σύμφωνα με το Σχήμα 2.8, τα περισσότερα άτομα και συγκεκριμένα το 91,1% εργάζονται υπό το καθεστώς της πλήρης απασχόλησης, ενώ το 8,9% υπό το καθεστώς της μερικής.



Σχήμα 2.8: Ραβδοδιάγραμμα με πλήθος ανά Είδος Απασχόλησης

Όπως παρατηρείται στον Πίνακα 2.16, το μεγαλύτερο ποσοστό τόσο των ανδρών, 95%, όσο και των γυναικών, 88%, εργάζεται με πλήρη απασχόληση. Περισσότερες είναι οι γυναίκες σε σχέση με τους άνδρες που εργάζονται σε καθεστώς μερικής απασχόλησης.

Είδος Απασχόλησης	**Άνδρας**		**Γυναίκα**	
	Σύνολο	Ποσοστό	Σύνολο	Ποσοστό
Πλήρης	34541	95%	24391	88%
Μερική	1833	5%	3320	12%

Πίνακας 2.16: Σύνολο και Ποσοστό ανδρών και γυναικών ανά Είδος Απασχόλησης

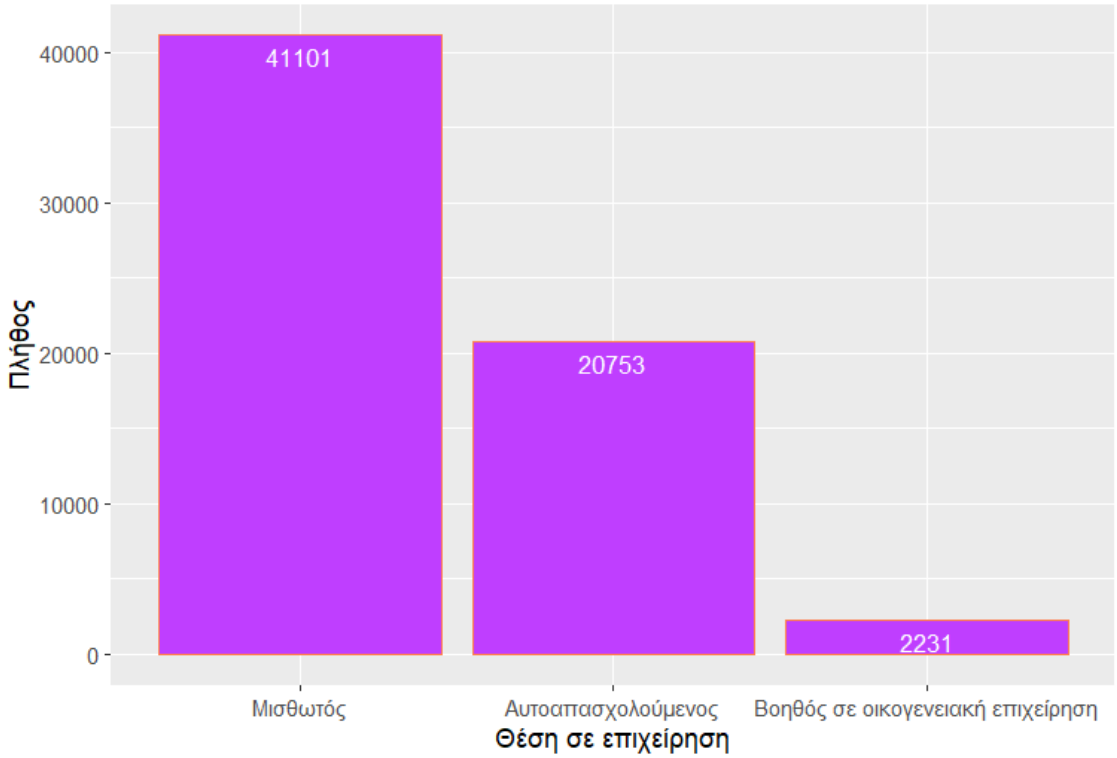
Εξετάζοντας το είδος της σύμβασης ανά την υπηκοότητα, παρατηρούμε στον Πίνακα 2.17 ότι το μεγαλύτερο ποστοστό ατόμων, ανεξαρτήτως υπηκοότητας, εργάζεται υπό το καθεστώς της πλήρης απασχόλησης. Τα άτομα με Ελληνική υπηκοότητα είναι περισσότερα και στις δύο κατηγορίες συμβάσεων.

Είδος Απασχόλησης	Ελλάδα		Ε.Ε.		Άλλη Χώρα	
	Σύνολο	Ποσοστό	Σύνολο	Ποσοστό	Σύνολο	Ποσοστό
Πλήρης	56941	92,1%	284	90,4%	1655	86,8%
Μερική	4867	7,9%	30	9,6%	252	13,2%

Πίνακας 2.17: Σύνολο και Ποσοστό ανά υπηκοότητα και ανά Είδος Απασχόλησης

2.2.2 Κύρια Θέση

Εξετάζοντας την κύρια θέση εργασίας στο σύνολο των εργαζομένων, παρατηρούμε από το Σχήμα 2.9 τα περισσότερα άτομα εργάζονται ως μισθωτοί, το 64,1%, με τους αυτοαπασχολούμενους να ακολουθούν, 32,4% και στην τελευταία θέση της απορρόφησης βρίσκονται οι βοηθοί σε οικογενειακή επιχείρησηνα αποτελούν το 3,5%.



Σχήμα 2.9: Ραβδοδιάγραμμα με πλήθος ανά Κύρια Θέση

Σύμφωνα με τον Πίνακα 2.18, ο οποίος παρουσιάζει τα αποτελέσματα της κατανομής των κύριων θέσεων ανα βαθμό αστικοποίησης, παρατηρούμε ότι η Περιφέρεια Πρωτευούσης αποτελείται στο μεγαλύτερο μέρος της από αυτοαπασχολούμενα άτομα, όπως και το Πολεοδεμικό Συγκρότημα της Θεσσαλονίκης και οι Αστικές περιοχές. Οι Αγρότικές περιοχές περιέχουν αρκετά υψηλό ποσοστό βοηθών σε οικογενειακές επιχειρήσεις, όπως και οι Ημιαστικές περιοχές.

Βαθμός αστικοποίησης	Μισθωτός	Αυταπασχολούμενος	Βοηθός
Περιφέρεια Πρωτευούσης	12,5%	29,2%	2,7%
ΠΣ Θεσσαλονίκης	3,2%	6,2%	1,6%
Αστική	23,6%	34,0%	14,2%
Ημιαστική	17,6%	13,3%	18,5%
Αγρότική	43,1%	17,3%	63,0%

Πίνακας 2.18: Ποσοστό ανά Κύρια Θέση ανά Βαθμό Αστικοποίησης

Επομένως, οι περισσότεροι μισθωτοί βρίσκονται σε αγρότικές περιοχές, οι αυτοαπασχολούμενοι σε αστικές περιοχές και οι βοηθοί σε οικογενειακές επιχειρήσεις σε αγρότικές περιοχές.

Τέλος, στον Πίνακα 2.19 και 2.20 εξετάζεται η θέση στην κύρια δραστηριότητα συνδυαστικά με το ανώτατο επίπεδο εκπαίδευσης. Παρατηρούμε ότι τα άτομα χωρίς στοιχειώδη εκπαίδευση απασχολούνται κυρίως ως μισθωτοί, τα άτομα με στοιχειώδη εκπαίδευση ως αυτοαπασχολούμενοι, οι απόφοιτοι γυμνασίου ως μισθωτοί, οι απόφοιτοι λυκείου ως μισθωτοί, οι απόφοιτοι ΙΕΚ ως μισθωτοί και οι κάτοχοι προπτυχιακού και μεταπτυχιακού τίτλου ως μισθωτοί. Οι περισσότερες θέσεις μισθωτών καλύπτονται από απόφοιτους λυκείου και κατόχους προπτυχιακού τίτλου, οι θέσεις αυτοαπασχολούμενων από απόφοιτους λυκείου, άτομα με στοιχειώδη εκπαίδευση και κατόχους προπτυχιακού τίτλου. Τέλος, οι θέσεις που αφορούν βοηθούς σε οικογενειακή επιχείρηση αποτελούνται κυρίως από απόφοιτους λυκείου και άτομα με στοιχειώδη εκπαίδευση, ενώ πολύ λίγες θέσεις καταλαμβάνουν τα άτομα που είναι κάτοχοι μεταπτυχιακού τίτλου.

Εκπαίδευση	Αυτοαπασχολούμενοι	Μισθωτοί
Χωρίς στοιχειώδη εκπαίδευση	98	125
Στοιχειώδη εκπαίδευση	4249	2386
Γυμνάσιο	2705	2826
Λύκειο	7370	14507
ΙΕΚ	1324	4813
Προπτυχιακό	4159	13092
Μεταπτυχιακό	848	3352

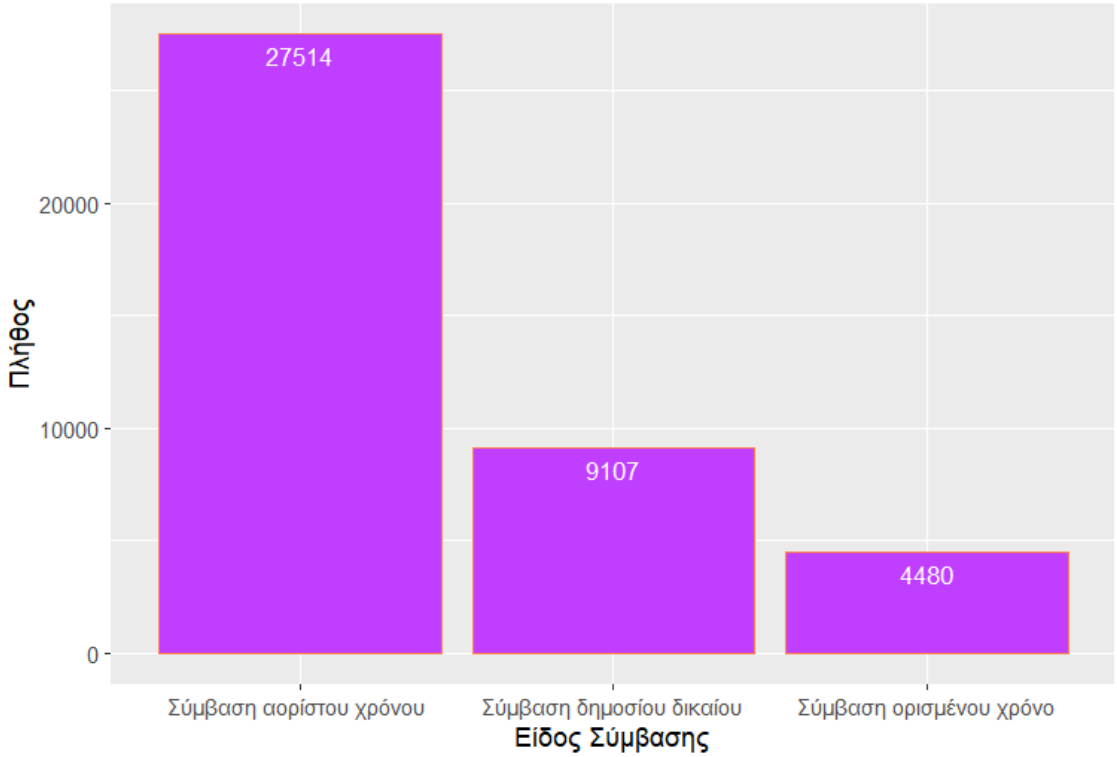
Πίνακας 2.19: Σύνολο ατόμων ανά κατηγορία απασχόλησης και Επίπεδο Εκπαίδευσης (Αυτοαπασχολούμενοι και Μισθωτοί)

Εκπαίδευση	Βοηθοί στην οικογενειακή επιχείρηση
Χωρίς στοιχειώδη εκπαίδευση	24
Στοιχειώδη εκπαίδευση	581
Γυμνάσιο	349
Λύκειο	928
ΙΕΚ	133
Προπτυχιακό	207
Μεταπτυχιακό	9

Πίνακας 2.20: Σύνολο ατόμων ανά κατηγορία απασχόλησης και Επίπεδο Εκπαίδευσης (Βοηθοί στην οικογενειακή επιχείρηση)

2.2.3 Είδος Σύμβασης

Το μεγαλύτερο ποσοστό των εργαζομένων απασχολείται με σύμβαση αορίστου χρόνου, το 67%, το 22,1% εργάζεται με μόνιμη σύμβαση δημοσίου δικαίου και τέλος, το 10,9% με προσωρινή σύμβαση ορισμένου χρόνου.



Σχήμα 2.10: Ραβδοδιάγραμμα με πλήθος ανά Είδος Σύμβασης

Από τον Πίνακα 2.21 και 2.22 παρατηρούμε ότι οι περισσότερες συμβάσεις που είναι μόνιμες και αφορούν θέσεις δημοσίου δικαίου βρίσκονται σε αστικές περιοχές και στην Περιφέρεια της Πρωτεύσης και οι

λιγότερες στο Πολεοδομικό Συγκρότημα της Θεσσαλονίκης. Οι περισσότερες συμβάσεις αορίστου χρόνου παρατηρούνται στην Περιφέρεια της Πρωτεούσης και σε αστικές περιοχές, ενώ οι λιγότερες στο Πολεοδομικό Συγκρότημα Θεσσαλονίκης. Τέλος, οι προσωρινές συμβάσεις ή ορισμένου χρόνου στη Περιφέρεια Πρωτεούσης και σε αστικές περιοχές, με τις λιγότερες να εμφανίζονται πάλι στο Πολεοδομικό Συγκρότημα Θεσσαλονίκης.

Βαθμός Αστικοποίησης	**Δημοσίου Δικαίου**	**Αορίστου χρόνου**
Περιφέρεια Πρωτευούσης	515	2024
ΠΣ Θεσσαλονίκης	126	426
Αστική	977	2001
Ημιαστική	337	809
Αγρότική	398	1048

Πίνακας 2.21: Σύνολο ατόμων ανά Είδος Σύμβασης και Βαθμό Αστικοποίησης (Δημοσίου Δικαίου και Αορίστου χρόνου)

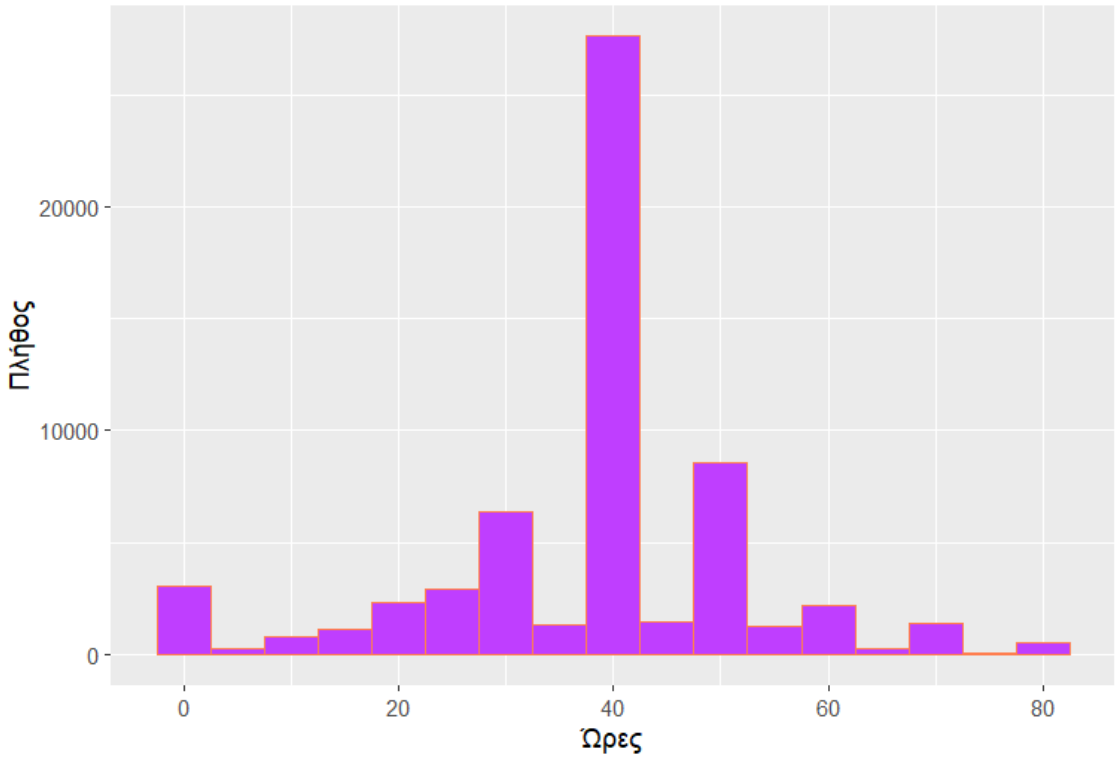
Βαθμός Αστικοποίησης	**Προσωρινή**
Περιφέρεια Πρωτευούσης	218
ΠΣ Θεσσαλονίκης	31
Αστική	232
Ημιαστική	114
Αγρότική	183

Πίνακας 2.22: Σύνολο ατόμων ανά Είδος Σύμβασης και Βαθμό Αστικοποίησης (Προσωρινή/Ορισμένου Χρόνου)

2.2.4 Εβδομαδιαίες Ώρες Εργασίας

Ο μέσος όρος των εβδομαδιαίων ωρών εργασίας για την Ελλάδα το 2022 ήταν 40.5 ώρες, όπως αναπαρίσταται και στο Σχήμα 2.11.

Συγκρίνοντας τις εβδομαδιαίες ώρες εργασίας της Ελλάδας με τις αντίστοιχες των χωρών της Ευρωπαϊκής Ένωσης, αλλά και της ευρύτερης περιοχής της Ευρώπης, όπως παρουσιάζονται από την Eurostat, προκύπτει ο Πίνακας 2.23. Ο μέσος όρος εργατικών ωρών ανά εβδομάδα για τις 27 χώρες της Ευρωπαϊκής Ένωσης είναι 37,3 ώρες, ενώ για την περιοχή της Ευρώπης οι 36,6 ώρες. Η Ελλάδα βρίσκεται στην 2η θέση, ενώ στην 1η θέση βρίσκεται η Σερβία με 43,2 ώρες.



Σχήμα 2.11: Ιστόγραμμα με μέσες εβδομαδιαίες ώρες εργασίας

Χώρα	**Ώρες Εργασίας**
Σερβία	43,2
Ελλάδα	40,5
Ευρωπαϊκή Ένωση	37,3
Ευρώπη	36,6

Πίνακας 2.23: Μέσες εβδομαδιαίες ώρες εργασίας στην Ευρώπη σύμφωνα με την Eurostat

Οι μέσες εβδομαδιαίες ώρες εργασίας για τα δύο φύλα παρατηρούμε ότι οι άνδρες εργάζονται 40,54 ώρες την εβδομάδα, ενώ οι γυναίκες 34,82 ώρες, Πίνακας 2.24 . Κάτι το οποίο δικαιολογείται από το γεγονός ότι περισσότερες γυναίκες, σε σχέση με άνδρες εργάζονται υπό το καθεστώς μερικής απασχόλησης.

Φύλο	**Ώρες Εργασίας**
Άνδρας	40,54
Γυναίκα	34,82

Πίνακας 2.24: Ο μέσος όρος εβδομαδιαίων ωρών εργασίας για τους άνδρες και τις γυναίκες

Εξετάζοντας πως διαμορφώνονται οι εβδομαδιαίες ώρες εργασίας ανά

βαθμό αστικοποίησης παρατηρούμε στον Πίνακα 2.25 ότι δεν υπάρχουν μεγάλες διαφορές στις μέσες εβδομαδιαίες ώρες εργασίας. Παρόλα αυτά φαίνεται σε ημιαστικές περιοχές τα άτομα να εργάζονται περισσότερες ώρες, ενώ σε αστικές περιοχές λιγότερες.

Βαθμός Αστικοποίησης	**Ώρες εργασίας**
Περιφέρεια Πρωτευούσης	37,51
ΠΣ Θεσσαλονίκης	38,64
Αστική	37,13
Ημιαστική	38,85
Αγρότική	38,31

Πίνακας 2.25: Πίνακας Βαθμού Αστικοποίησης και Μέσων Εβδομαδιαίων Ωρών Εργασίας

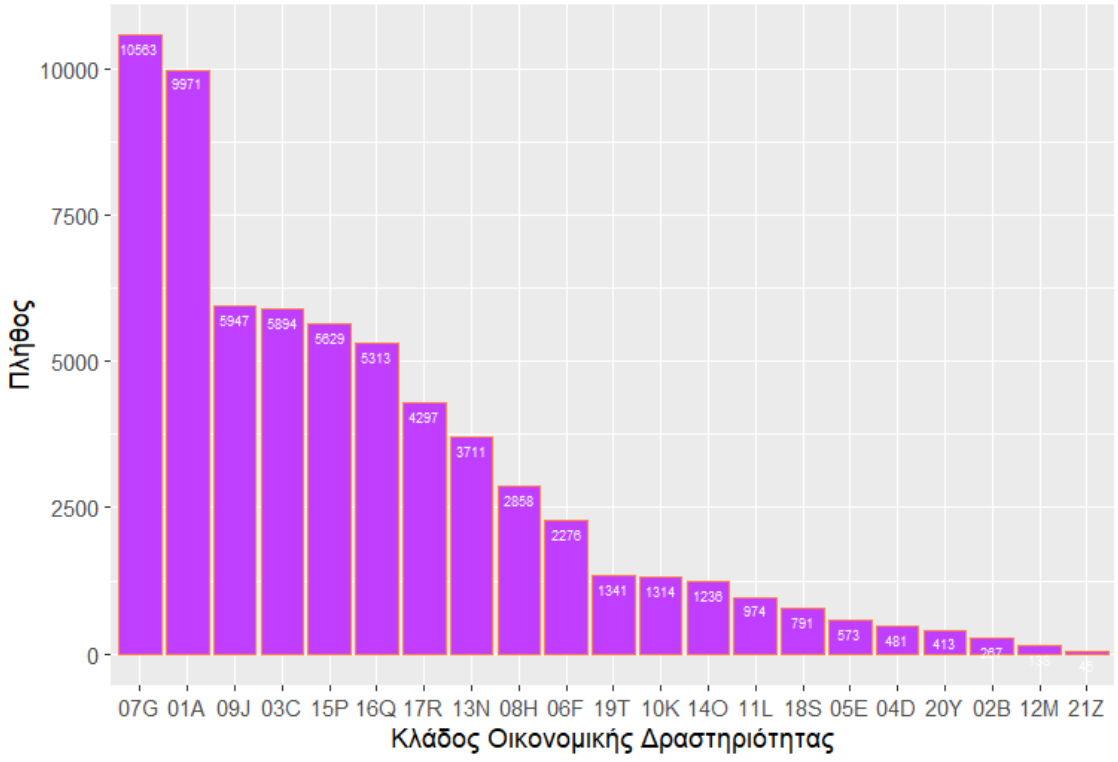
2.2.5 Κλάδος Οικονομικής Δραστηριότητας-Κύρια Δραστηριότητα

Όσον αφορά τον κλάδο οικονομικής δραστηριότητας, Πίνακας 2.26 και Σχήμα 2.12, το μεγαλύτερο ποσοστό των εργαζομένων απασχολείται στο Χονδρικό και Λιανικό Εμπόριο, ακολουθεί ο τομέας της Γεωργίας/Αλιείας/Δασοκομίας, η εστίαση και η παροχή καταλυμάτων. Το μικρότερο ποσοστό απασχόλησης παρουσιάζεται σε Δραστηριότητες Νοικοκυριών, στα Ορυχεία/Λατομεία, την Διαχείρισης Ακίνητης Περιουσίας και τέλος, η Δραστηριότητα Ετερόδικων Οργανισμών και φορέων.

Ενώ αν εστιάσουμε στην κύρια δραστηριότητα των εργαζομένων το 21,3% απασχολείται σαν πωλητής ή σαν πάροχος υπηρεσιών και το 20,5% σαν επαγγελματίας. Τα ποσοστά αυτά δικαιολογούν και τις υψηλές θέσεις σε μισθωτούς και αυτοαπασχολούμενους. Στην τελευταία θέση, με ποσοστό 3,1% βρίσκονται τα Ανώτερα Διευυντικά και Διοικητικά Στελέχη.

Κλάδος Οικονομικής Δραστηριότητας	Σύνολο	Ποσοστό
07Γ = Χονδρικό-Λιανικό εμπόριο	10563	25,6%
01Α = Γεωργία/Αλεία/Δασοκμία	9971	24,2%
09Θ = Εστίαση/Παροχή καταλυμάτων	5947	14,4%
03Σ = Μεταποίηση	5894	14,3%
15Π = Δημόσια Διοίκηση και Άμυνα	5629	13,6%
18Χ = Εκπαίδευση	5313	12,9%
17Ρ = Ανθρ.Υγεία και Κοιν. Ευημερία	4297	10,4%
13Ν = Επαγ., Επιστ. και Τεχνικές Δραστηριότητες	3711	9%
08Η = Μεταφορά και Αποθήκευση	2858	6,9%
06Φ = Κατασκευές	2276	5,5%
19Τ= Άλλες Δραστ. Παροχής Υπηρεσιών	1341	3,2%
10Κ = Ενημέρωση και Επικοινωνία	1314	3,1%
14Ο = Διοικητικές και Υποστηρικτές δραστ.	1236	3%
11Λ = Χρηματοπιστωτικές και Ασφαλιστικές δραστηριότητες	974	2,3%
18Σ = Τέχνες/Διασκέδαση/Ψυχαγωγία	791	1,9%
05Ε = Παροχή νερού/Επ.Λυμάτων/Διαχ.Αποβλήτων	573	1,3%
04Δ = Παροχή ρεύματος, φυσικού αερίου	481	1,1%
20Ψ = Δραστηριότητες νοικοκυριών	413	1%
02Β = Ορυχεία/Λατομεία	267	0,6%
12Μ = ιαχείριση ακίνητης περιουσίας	138	0,3%
21Ζ = Δραστ. Ετερόδικων οργανισμών και φορέων	45	0,1%

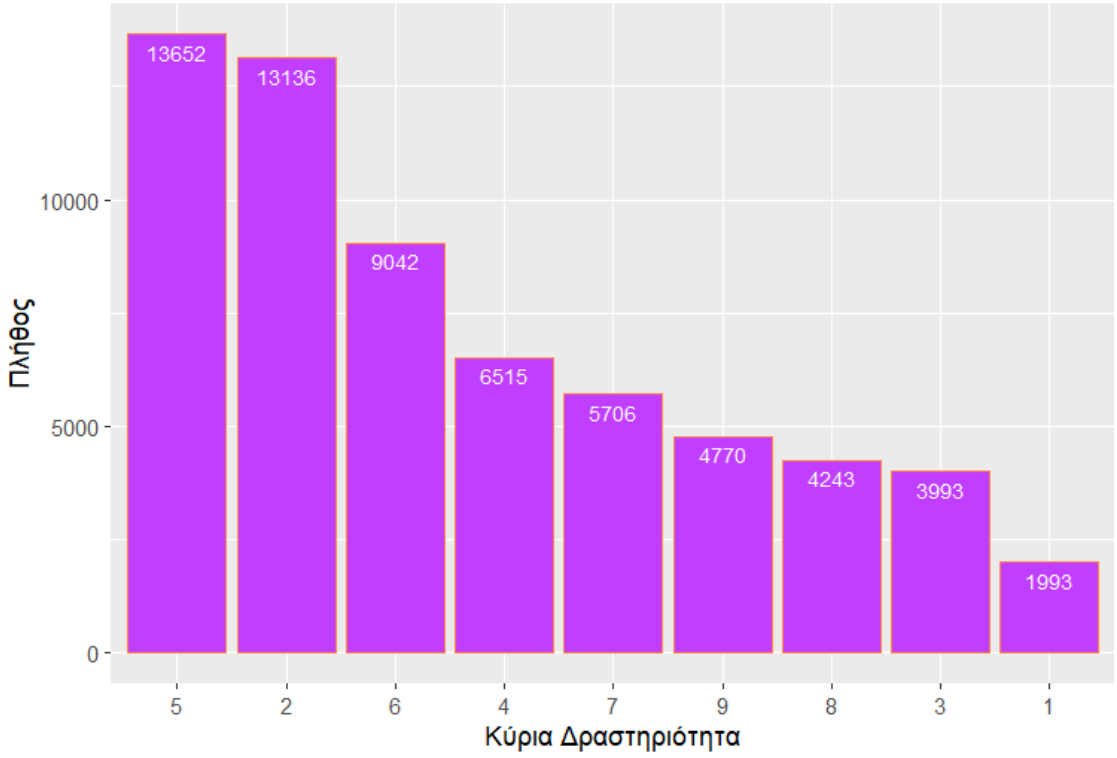
Πίνακας 2.26: Σύνολο και Ποσοστό ανά Κλάδο Οικονομικής Δραστηριότητας



Σχήμα 2.12: Ραβδοδιάγραμμα με πλήθος ανά Κλάδο Οικονομικής Δραστηριότητας

Κύρια Δραστηριότητα	Σύνολο	Ποσοστό
5 = Παροχή Υπηρεσιών και Πωλητές	13652	21,3%
2 = Επαγγελματίες	13136	20,5%
6 = Ειδ. γεωργοί/Κτηνοτρόφοι/Δασοκόμοι/Αλιείς	9042	14,1%
4 = Υπάλληλοι γραφείου	6515	10,1%
7 = Ειδικευόμενοι Τεχνίτες	5706	8,9%
9 = Ανειδίκευτοι εργάτες/Χειρωνάκτες	4770	7,4%
8 = Χειριστές Βιομηχανικών Μηχανημάτων	4243	6,6%
3 = Τεχνικοί και συναφή ασκούντες	3993	6,2%
1 = Ανώτερα Διευθυντικά και Διοικητικά Στελέχη	1993	3,1%

Πίνακας 2.27: Σύνολο και Ποσοστό ανά Κύρια Δραστηριότητα



Σχήμα 2.13: Ραβδοδιάγραμμα με πλήθος ανά Κύρια Δραστηριότητα

Στον Πίνακα 2.28 παρατηρούμε ότι το μεγαλύτερο ποσοστό των εργαζομένων γυναικών απασχολείται ως επαγγελματίες και στην παροχή υπηρεσιών ή τις πωλήσεις, ομοίως και οι άνδρες. Η μικρότερη απορρόφηση των γυναικών παρουσιάζεται σε επαγγέλματα που αφορούν ειδικευόμενους τεχνίτες και επαγγέλματα που αφορούν τον χειρισμό βιομηχανικών μηχανημάτων, ακόμα και σε θέσεις που αφορούν ανώτερα διευθυντικά και διοικητικά στελέχη. Ενώ η μικρότερη απορρόφηση των ανδρών παρουσιάζεται σε τομείς που αφορούν ανώτερα διευθυντικά και διοικητικά στελέχη.

Κύρια Δραστηριότητα	Άνδρας		Γυναίκα	
	Σύνολο	Ποσοστό	Σύνολο	Ποσοστό
Ανώτερα διευθυντικά και διοικητικά στελέχη	1366	3,9%	627	2,3%
Επαγγελματίες	5721	16.1%	7415	26.9%
Τεχνικοί και συναφή ασκούντες	2285	6.4%	1708	6.2%
Υπάλληλοι γραφείου	2576	7.3%	3939	14.3%
Παροχή υπηρεσιών/πωλητές	6921	19.5%	6731	24.4%
Ειδ. γεωργοί/κτηνοτρόφοι/δασοκόμοι/αλιείς	5365	15,1%	3677	13,3 %
Ειδικευμένοι τεχνίτες	5216	14,7%	490	1,8%
Χειριστές βιομηχανικών εγκαταστάσεων	3846	10,9%	397	1,4%
Ανειδίκευτοι εργάτες,/χειρωνάκτες	2164	6,1%	2606	9,4 %

Πίνακας 2.28: Σύνολο και Ποσοστό ατόμων ανά Κύρια Δραστηριότητα με βάση το Φύλο

Τα παραπάνω αποτελέσματα, συνδυαστικά με τον Πίνακα 2.8 και Σχήμα 2.9 όπου οι περισσότεροι εργαζόμενοι άνδρες διανέμουν σε αγρότικές περιοχές δικαιολογεί και την αυξημένη τους απορρόφηση σε χειρωνακτικά επαγγέλματα, όπως η γεωργία, η κτηνοτροφία, η αλιεία κ.λπ..

Στους πίνακες που ακολουθούν εξετάζεται η απορρόφηση των ατόμων σε τομείς κύριας δραστηριότητας ανάλογα με το ανώτατο επιπεδο εκπαίδευσης. Στον Πίνακα 2.29 παρουσιάζεται η απορρόφηση των ατόμων χωρίς στοιχειώδη εκπαίδευση, όπου τα περισσότερα άτομα εργάζονται σε θέσεις που αφορούν χειρωνακτικά επαγγέλματα, δηλαδή ως κτηνοτρόφοι, ειδικευόμενοι γεωργοί, δασοκόμοι και αλιείς και ως ανειδίκευτοι εργάτες ή χειρωνάκτες.

Κύρια Δραστηριότητα	Χωρίς στοιχειώδη εκπαίδευση
Ανώτερα διευθυντικά και διοικητικά στελέχη	0
Επαγγελματίες	0
Τεχνικοί και συναφή ασκούντες	0
Υπάλληλοι γραφείου	1
Παροχή υπηρεσιών/πωλητές	31
Ειδ.γεωργοί/κτηνοτρόφοι/δασοκόμοι/αλιείς	92
Ειδικευμένοι τεχνίτες	23
Χειριστές βιομηχανικών εγκαταστάσεων	10
Ανειδίκευτοι εργάτες/ειρωνάκτες	90

Πίνακας 2.29: Σύνολο ατόμων ανά Κύρια Δραστηριότητα και Επίπεδο Εκπαίδευσης (Χωρίς στοιχειώδη εκπαίδευση)

Στον Πίνακα 2.30 παρουσιάζεται η απορρόφηση των ατόμων με στοιχειώδη εκπαίδευση όπου τα περισσότερα άτομα εργάζονται ως ειδικευόμενοι γεωργοί, κτηνοτρόφοι, δασοκόμοι και αλιείς και ως ανειδίκευτοι εργάτες

ή χειρωνάκτες και ακολουθούν επαγγέλματα που αφορούν ειδικευόμενους τεχνίτες. Παρτηρούμε ότι η μεγαλύτερη απορρόφηση πραγματοποιείται σε ίδιες θέσεις με τα άτομα χωρίς στοιχειώδη εκπαίδευση, ενώ εισέρχονται και θέσεις με πιο ειδικευόμενα επαγγέλματα.

Κύρια Δραστηριότητα	**Στοιχειώδη εκπαίδευση**
Ανώτερα διευθυντικά και διοικητικά στελέχη	76
Επαγγελματίες	16
Τεχνικοί και συναφή ασκούντες συναφή	28
Υπάλληλοι γραφείου	110
Παροχή υπηρεσιών/πωλητές	892
Ειδ.γεωργοί/κτηνοτρόφοι/δασοκόμοι/αλιείς	3465
Ειδικευμένοι τεχνίτες	926
Χειριστές βιομηχανικών εγκαταστάσεων	458
Ανειδίκευτοι εργάτες/χειρωνάκτες	1242

Πίνακας 2.30: Σύνολο ατόμων ανά Κύρια Δραστηριότητα και Επίπεδο Εκπαίδευσης (Στοιχειώδη εκπαίδευση)

Στον Πίνακα 2.31 παρουσιάζεται η απορρόφηση των ατόμων που έχουν ολοκληρώσει το γυμνάσιο, λύκειο ή έχουν αποφοιτήσει από κάποιο ΙΕΚ. Τα άτομα που έχουν ολοκληρώσει το γυμνάσιο, δηλαδή την υποχρεωτική εκπαίδευση στην Ελλάδα, εργάζονται κυρίως ως ειδικευόμενοι γεωργοί, κτηνοτρόφοι, δασοκόμοι και αλιείς και στην παροχή υπηρεσιών ή ως πωλητές. Τα άτομα που είναι απόφοιτοι λυκείου εργάζονται κυρίως στην παροχή υπηρεσιών και στις πωλήσεις και τέλος τα άτομα που έχουν αποφοιτήσει από ΙΕΚ εργάζονται στην παροχή υπηρεσιών, στις πωλήσεις και ως υπάλληλοι γραφείου.

Κύρια Δραστηριότητα	**Γυμνάσιο**	**Λύκειο**	**ΙΕΚ**
Ανώτερα διευθυντικά και διοικητικά στελέχη	101	683	162
Επαγγελματίες	24	564	816
Τεχνικοί και συναφή ασκούντες	71	924	801
Υπάλληλοι γραφείου	196	2836	922
Παροχή υπηρεσιών/πωλητές	1062	7152	2079
Ειδ.γεωργοί/κτηνοτρόφοι/δασοκόμοι/αλιείς	1803	2989	260
Ειδικευμένοι τεχνίτες	930	2760	562
Χειριστές βιομηχανικών εγκαταστάσεων	759	2399	299
Ανειδίκευτοι εργάτες/χειρωνάκτες	922	2010	238

Πίνακας 2.31: Σύνολο ατόμων ανά Κύρια Δραστηριότητα και Επίπεδο Εκπαίδευσης (Γυμνάσιο, Λύκειο, ΙΕΚ)

Στον Πίνακα 2.32 παρατηρούμε ότι οι κάτοχοι προπτυχιακού τίτλου εργάζονται κυρίως ως επαγγελματίες, ακολουθούν οι παροχή υπηρεσιών

και οι πωλήσεις, καθώς και οι υπάλληλοι γραφείου. Τέλος, τα άτομα που κατέχουν μεταπτυχιακό τίτλο εργάζονται ως επαγγελματίες.

Κύρια Δραστηριότητα	**Προπτυχιακό**	**Μεταπτυχιακό**
Ανώτερα διευθυντικά και διοικητικά στελέχη	606	365
Επαγγελματίες	8731	2985
Τεχνικοί και συναφή ασκούντες	1850	319
Υπάλληλοι γραφείου	2142	308
Παροχή υπηρεσιών/πωλητές	2279	157
Ειδ.γεωργοί/κτηνοτρόφοι/δασοκόμοι/αλιείς	425	8
Ειδικευμένοι τεχνίτες	494	11
Χειριστές βιομηχανικών εγκαταστάσεων	312	6
Ανειδίκευτοι εργάτες/χειρωνάκτες	256	12

Πίνακας 2.32: Σύνολο ατόμων ανά Κύρια Δραστηριότητα και Επίπεδο Εκπαίδευσης (Προπτυχιακό, Μεταπτυχιακό)

Παρατηρώντας και τους 4 πίνακες οι θέσεις που αφορούν ανώτερα διευθυντικά και διοικητικά στελέχη αποτελούνται κυρίως από απόφοιτους λυκείου, άτομα με προπτυχιακό και μεταπτυχιακό τίτλο και καθόλου από άτομα χωρίς στοιχειώδη ή στοιχειώση εκπαίδευση. Παρατηρούμε ότι όσο ανεβαίνουν τα επίπεδα εκπαίδευσης τα άτομα απορροφώνται σε θέσεις που αφορούν μη χειρωνακτικά επαγγέλματα, σε αντίθεση με τα άτομα που βρίσκονται στα πρωταρχικά επίπεδα εκπαίδευσης. Όπως είδαμε επίσης ότι η απορρόφηση των γυναικών σε τέτοια επαγγέλματα είναι χαμηλή.

2.3 Σχετικά με τους Ανέργους

Σύμφωνα με τον Πίνακα 2.33 το 41,6% των ανέργων είναι απόφοιτοι λυκείου και το 29,2% είναι κάτοχοι προπτυχιακού τίτλου. Το μικρότερο ποσοστό καταλαμβάνουν τα άτομα χωρίς στοιχειώδη εκπαίδευση και οι κάτοιχοι μεταπτυχιακού τίτλου.

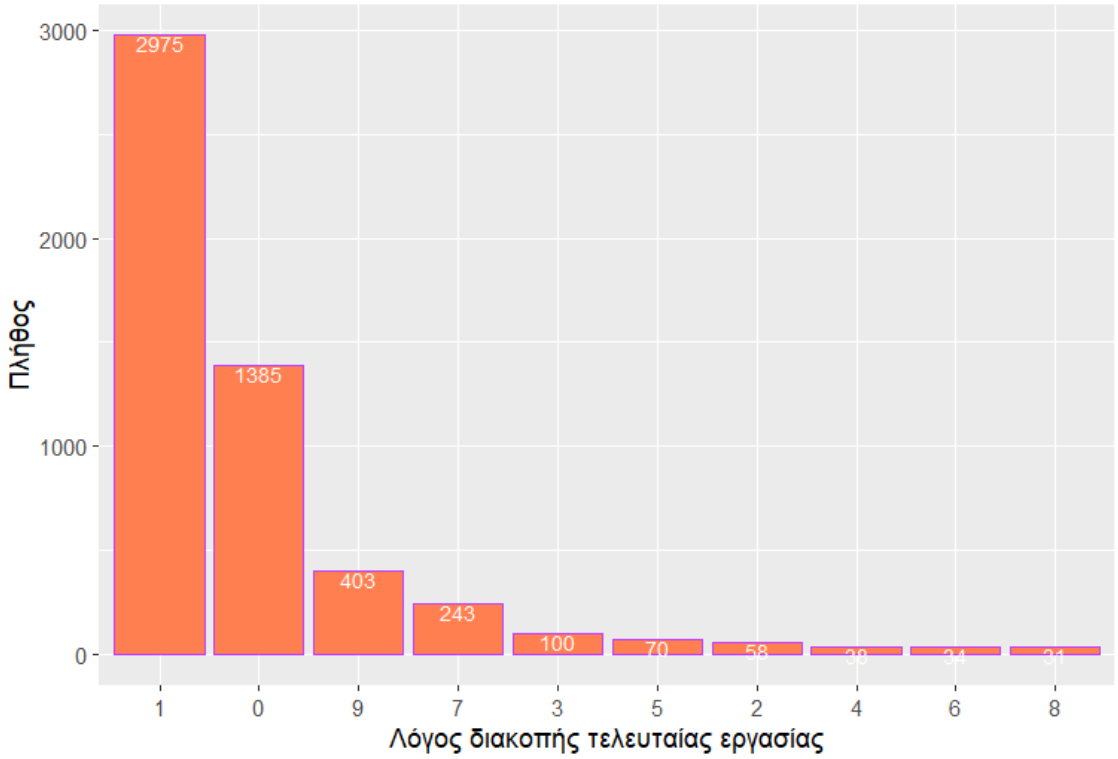
Εκπαίδευση	**Σύνολο**	**Ποσοστό**
Χωρίς στοιχειώδη εκπαίδευση	73	0.9%
Στοιχειώδη εκπαίδευση	1025	12,0%
Γυμνάσιο	967	11,4%
Λύκειο	3540	41,6%
ΙΕΚ	1032	12,1%
Προπτυχιακό	1637	19,2%
Μεταπτυχιακό	236	2,8%

Πίνακας 2.33: Σύνολο ατόμων και ποσοστά ανά Επίπεδο Εκπαίδευσης

Ο κυριότερος λόγος διακοπής της τελευταίας εργασίας, όπως φαίνεται στον Πίνακα 2.34 και Σχήμα 2.14, αποτελεί η περιορισμένη διάρκεια της σύμβασης, με ποσοστό 55,8% στο σύνολο των ανέργων και ακολουθεί η απόλυση ή η διακοπή λειτουργίας της επιχείρησης με 26%. Επομένως, τα περισσότερα άτομα είναι άνεργα διότι εργάστηκαν με προσωρινή σύμβαση.

Λόγος διακοπής τελευταίας εργασίας	**Σύνολο**	**Ποσοστό**
Η εργασία ήταν περιορισμένης διάρκειας και τελείωσε	2975	55,8%
Διότι απολύθηκε ή έκλεισε η επιχείρηση	1385	26%
Για άλλους λόγους	403	7,5%
Άλλοι προσωπικοί λόγοι	243	4,5%
Για άλλους οικογενειακούς λόγους	100	1,9%
Λόγω ασθένειας ή ανικανότητας	70	1,3%
Φροντίζει μικρά παιδιά ή εξαρτώμενους ενήλικες	58	1,1%
Λόγω συνταξιοδότησης	38	0,8%
Λόγω εκπαίδευσης ή επιμόρφωσης	34	0,6%
Διότι στρατεύτηκε	31	0,5%

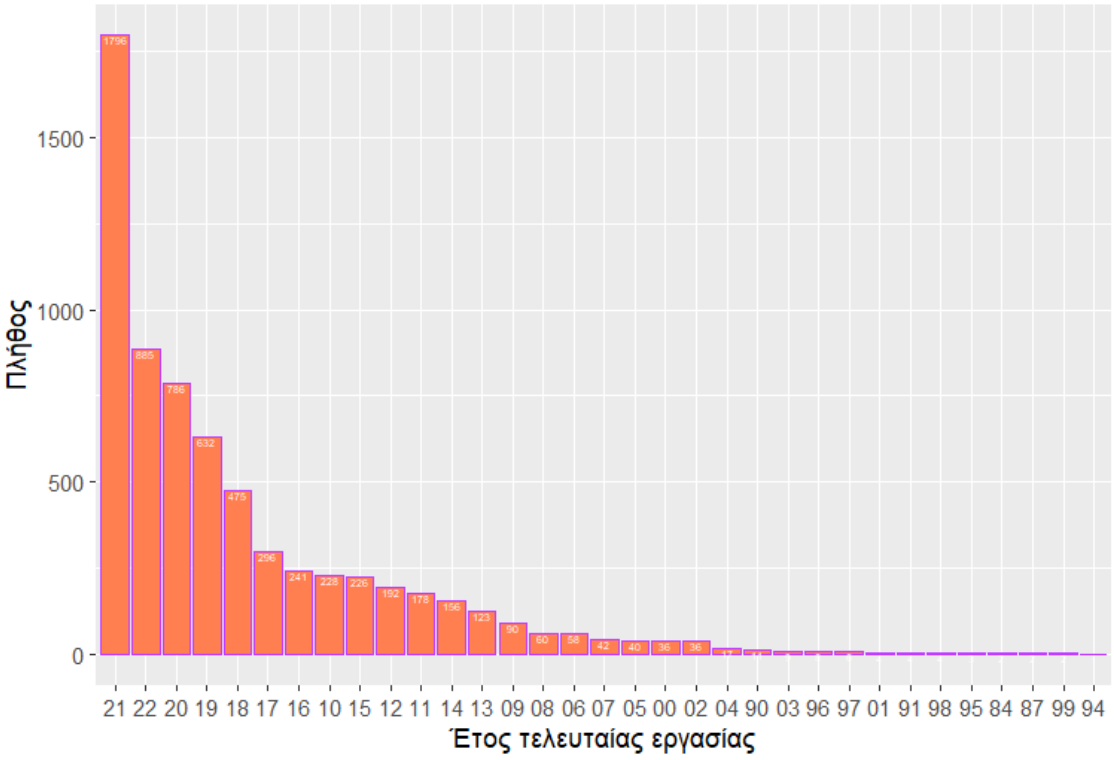
Πίνακας 2.34: Σύνολο και Ποσοστό ανά λόγο διακοπής τελευταίας εργασίας



Σχήμα 2.14: Ραβδοδιάγραμμα με πλήθος ανά λόγο διακοπής τελευταίας εργασίας

Όσον αφορά το έτος τελευταίας εργασίας, Σχήμα 2.15, το 27,0% είναι άνεργο από το 2021, ενώ το 13,3% σταμάτησε να εργάζεται το ίδιο έτος

της έρευνας, το 2022. Υπάρχουν και μικρά ποσοστά μακροχρόνιων ανέργων από την δεκαετία του '90, καθώς και απο τις αρχές του 2000.



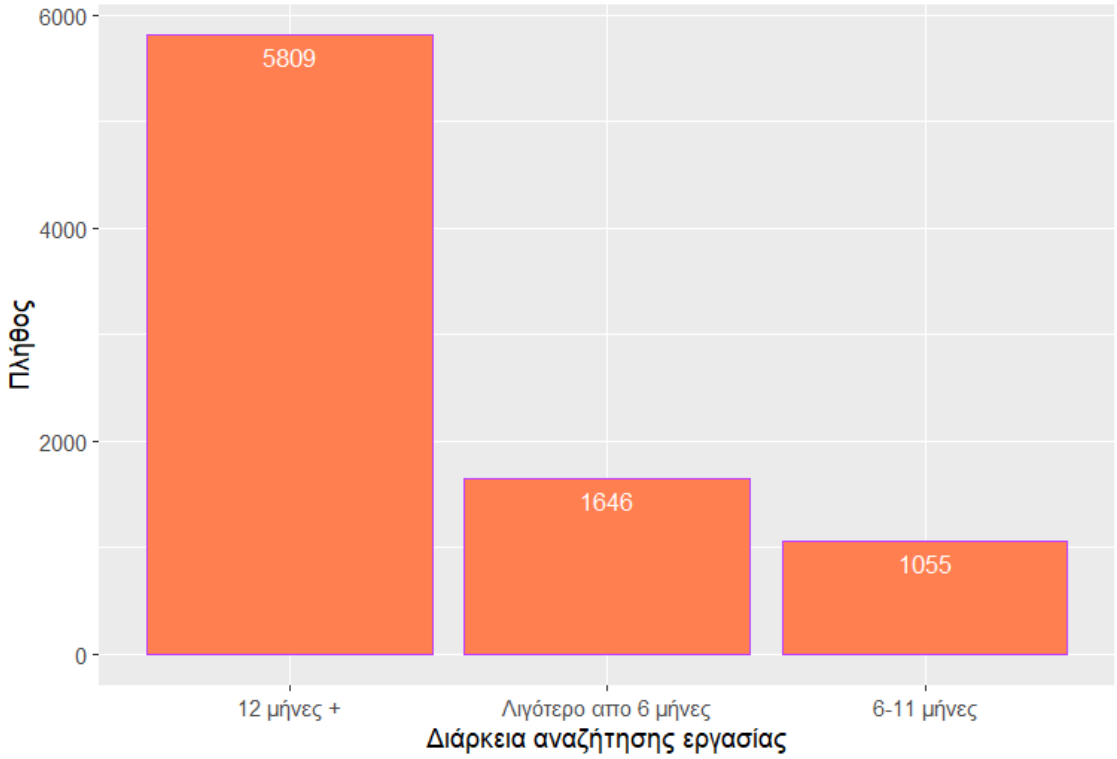
Σχήμα 2.15: Ραβδοδιάγραμμα με πλήθος ανά έτος τελευταίας εργασίας

Εξετάζοντας τον λόγο διακοπής της τελευταίας εργασίας ανάμεσα στα δύο φύλα στον Πίνακα 2.35 παρατηρούμε ότι οι περισσότερες γυναίκες σταματούν να εργάζονται λόγω περιορισμένης διάρκειας της εργασίας, ομοίως και οι άνδρες. Ακολουθεί ο λόγος της απόλυσης ή κλεισίματος μιας επιχείρησης για οικονομικούς λόγους με σχεδόν ίδια ποσότητα ατόμων και στα δύο φύλα, με τις γυναίκες να είναι ελάχιστα περισσότερες. Για λόγους που αφορούν την φροντίδα μικρών παιδιών ή εξαρτημένων ηλικιωμένων ατόμων, ή ακόμα και άλλιων οικογενειακών λόγων οι γυναίκες είναι περισσότερες, σε σχέση με τους άνδρες. Ενώ το μικρότερο ποσοστό των γυναικών είναι λόγω στράτευσης.

Λόγος Διακοπής τελευταίας εργασίας	Άνδρας	Γυναίκα
Διότι απολύθηκε ή έκλεισε η επιχείρηση	678	707
Η εργασία ήταν περιορισμένης διάρκειας και τελείωσε	1294	1681
Φροντίζει μικρά παιδιά ή εξαρτώμενους ενήλικες	1	57
Για άλλους οικογενειακούς λόγους	23	77
Λόγω εκπαίδευσης ή επιμόρφωσης	22	12
Λόγω ασθένειας ή ανικανότητας	33	37
Λόγω συνταξιοδότησης	28	10
Άλλοι προσωπικοί λόγοι	96	147
Διότι στρατεύτηκε	27	4
Για άλλους λόγους	187	216

Πίνακας 2.35: Λόγος διακοπής τελευταίας εργασίας ανά Φύλο

Παρατηρώντας το Σχήμα 2.16 το 68,2% των ανέργων αναζητούν εργασία για περισσότερο από 12 μήνες. Πάνω από τους μισούς ανέργους αναζητούν εργασία για πάνω από ένα χρόνο, κάτι το οποίο επιβεβαιώνεται και από τα νούμερα των ατόμων στο έτος της τελευταίας εργασίας. Το 19,3% αναζητά εργασία για λιγότερο από 6 μήνες και το 12,5% για διάστημα μεταξύ 6 με 11 μηνών.



Σχήμα 2.16: Ραβδοδιάγραμμα με πλήθος ανά διάρκεια αναζήτησης εργασίας

Σύμφωνα με τον Πίνακα 2.36 το μεγαλύτερο ποσοστό των ανδρών, 66%, αναζητά εργασία για πάνω από 12 μήνες, ομοίως το 70% των γυναικών.

Επομένως, τα δύο φύλα χρειάζονται σχεδόν το ίδιο χρονικό διάστημα για να βρούν την επόμενη εργασία τους.

Φύλο	**12 μήνες +**	**Λιγότερο από 6 μήνες**	**6-11 μήνες**
Άνδρας	66%	21%	13%
Γυναίκα	70%	18%	12%

Πίνακας 2.36: Ποσοστό ατόμων ανά φύλο ανά διάρκεια αναζήτησης εργασίας

Σύμφωνα με τον Πίνακα 2.37, ο οποίος εξετάζει την διάρκεια αναζήτησης εργασίας ανά ανώτατο επίπεδο εκπαίδευσης, παρατηρούμε ότι το μεγαλύτερο ποσοστό ατόμων σε καθένα από τα 7 επίπεδα εκπαίδευσης αναζητά εργασία μεταξύ 6 και 11 μηνών, ενώ τα αμέσως επόμενα μεγαλύτερα ποσοτά εμφανίζονται για διάρκεια μεγαλύτερη των 12 μηνών.

Επίπεδο Εκπαίδευσης	**Λιγότερο από 6 μήνες**	**6-11 μήνες**	**12 μήνες+**
Χωρίς στοιχειώδη εκπαίδευση	4,6%	1,6%	93,8%
Στοιχειώδη εκπαίδευση	18,3%	10,2%	71,5%
Γυμνάσιο	19,9%	10,1%	70,0%
Λύκειο	18,0%	13,0%	69,0%
ΙΚΕ	22,1%	12,5%	65,4%
Προπτυχιακό	20,1%	14,3%	65,6%
Μεταπτυχιακό	28,8%	13,1%	58,1%

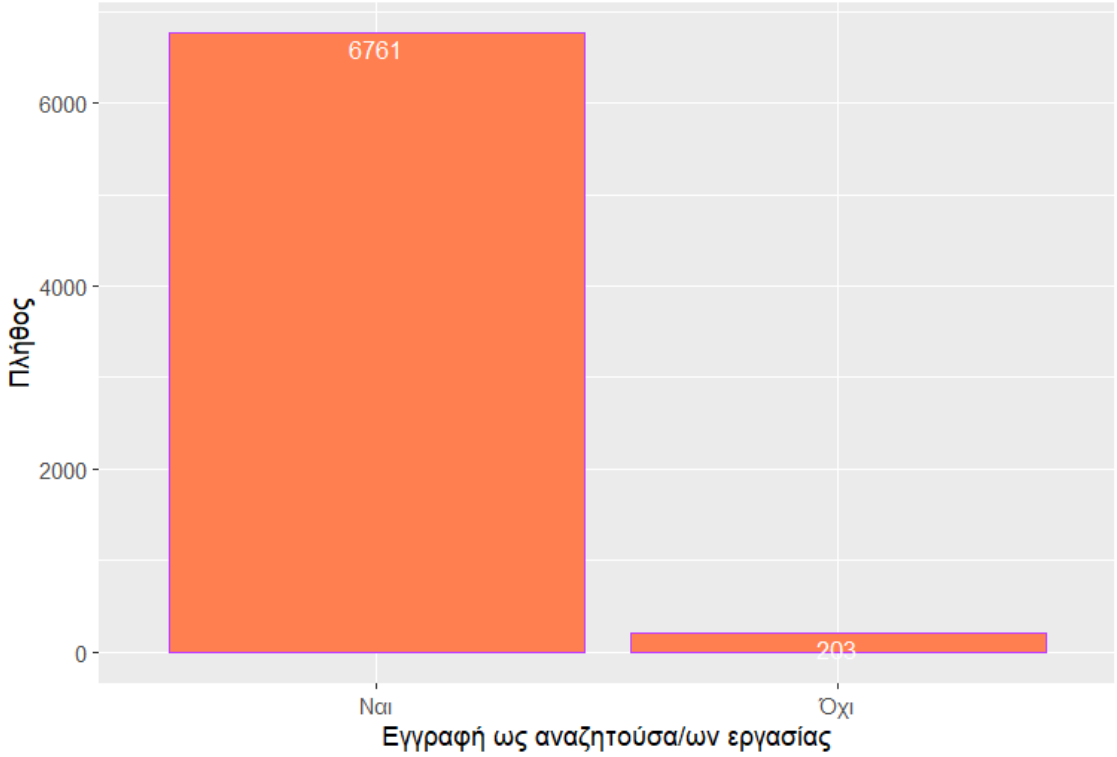
Πίνακας 2.37: Ποσοστό ατόμων διάρκειας αναζήτησης εργασίας ανά Ανώτατο Επίπεδο Εκπαίδευσης

Εξετάζοντας τον Πίνακα 2.33, σύμφωνα με τον οποίο οι περισσότεροι άνεργοι είναι απόφοιτοι της δευτεροβάθμιας εκπαίδευσης (λύκειο) και ακολουθούν οι κάτοχοι προπτυχιακού τίτλου, συνδυαστικά με τον 2.37 τα άτομα που αναζητούν περισσότερο εργασία είναι οι απόφοιτοι λυκείου και οι κάτοιχοι προπτυχιακού τίτλου.

Από το Σχήμα 2.17 παρατηρούμε ότι το 97,1% των ανέργων έχει γραφτεί σε κάποιο πρόγραμμα ως αναζητών εργασίας, ενώ το 82,5% λαμβάνει κάποιο επίδομα μέσω προγράμματος της Δημόσιας Υπηρεσίας Απασχόλησης (ΔΥΠΑ) (πρώην ΟΑΕΔ), Σχήμα 2.18. Σύμφωνα με τον Πίνακα 2.38, από τα άτομα που έχουν εγγραφεί σε κάποιο πρόγραμμα αναζήτησης εργασίας το μεγαλύτερο μέρος τους αποτελείται από γυναίκες, που αποτελούν το 59,3%.

Φύλο	Ναι	Όχι
Άνδρας	2754	81
Γυναίκα	4007	122

Πίνακας 2.38: Πίνακας με εγγραφή ως αναζητούσα/ων εργασίας ανα Φύλο

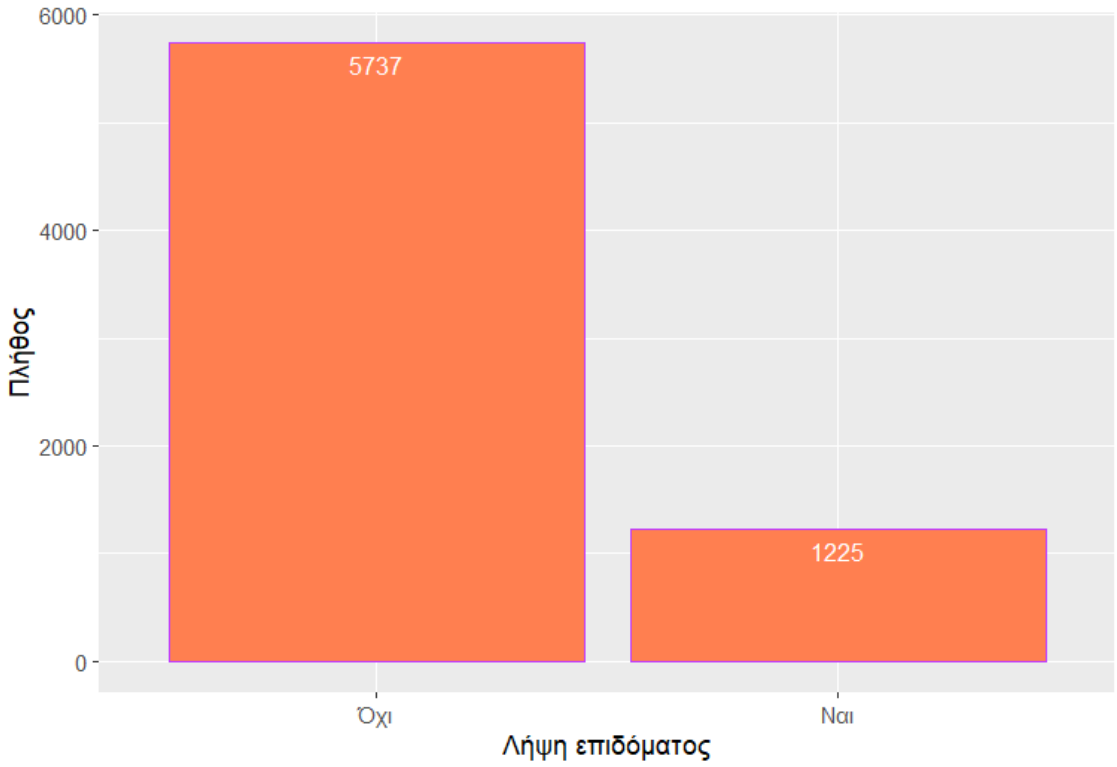


Σχήμα 2.17: Ραβδοδιάγραμμα με πλήθος εγγεγραμμένων ως αναζητούσα/ων εργασίας

Από τα άτομα που λαμβάνουν επίδομα το μεγαλύτερο ποσοστό καταλαμβάνουν οι γυναίκες, σύμφωνα με τον Πίνακα 2.39.

Φύλο	Ναι	Όχι
Άνδρας	545	2285
Γυναίκα	680	3452

Πίνακας 2.39: Πίνακας με λήψη επιδόματος ή όχι ανα Φύλο



Σχήμα 2.18: Ραβδοδιάγραμμα με πλήθος λήψης ή όχι επιδόματος

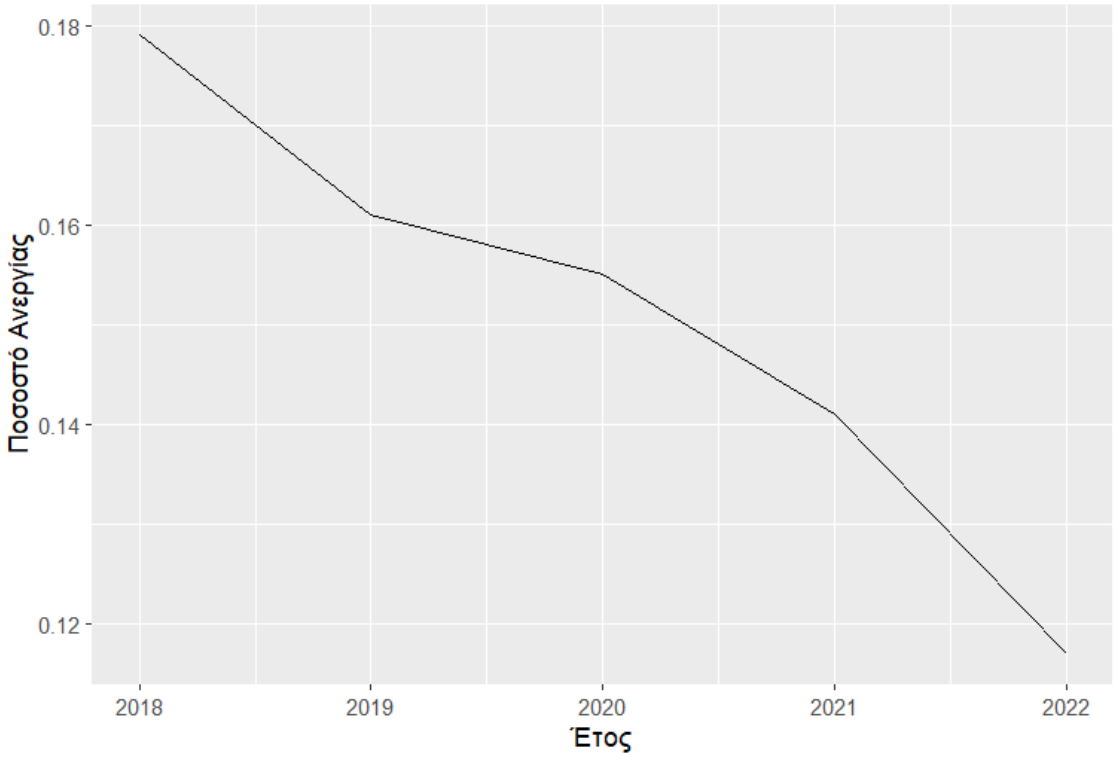
2.4 Η Πορεία της Ανεργίας από το 2018-2022

Στη παρούσα παράγραφο, ελέγχεται η εξέλιξη της ανεργίας σε κάποιες από τις κατηγορίες που εξετάστηκε και στην παράγραφο 2.1. Για αυτό το σκοπό, τα δεδομένα του κάθε έτους αναλύονται ξεχωριστά και στο τέλος δημιουργείται ένας πίνακας που περιέρχει τα συγκεντρωτικά δεδομένα, ο οποίος αναλύεται εκ νέου και προκύπτουν τα ακόλουθα αποτελέσματα και διαγράμματα, με την χρήση του ggplot2. Τα ποσοστά της ανεργίας έχουν υπολογιστεί για την κάθε κατηγορία, σύμφωνα με τον ορισμόυ του Διεθνούς Οργανισμού Εργασίας, ως το ποσοστό των ανέργων προς το σύνολο των ανέργων και των εργαζομένων.

Παρατήρειται στον Πίνακα 2.10 και στο Σχήμα 2.11, τα πέντε τελευταία έτη, από το 2018 έως και το 2022, το ποσοστό της ανεργίας παρουσιάζει σταδιακή μείωση, με το 2022 να έχει το μικρότερο ποσοστό.

Έτος	Ποσοστό ανεργίας
2018	17,9%
2019	16,1%
2020	15,5%
2021	14,1%
2022	11,7%

Πίνακας 2.40: Ποσοστά ανεργίας από το 2018 έως το 2022



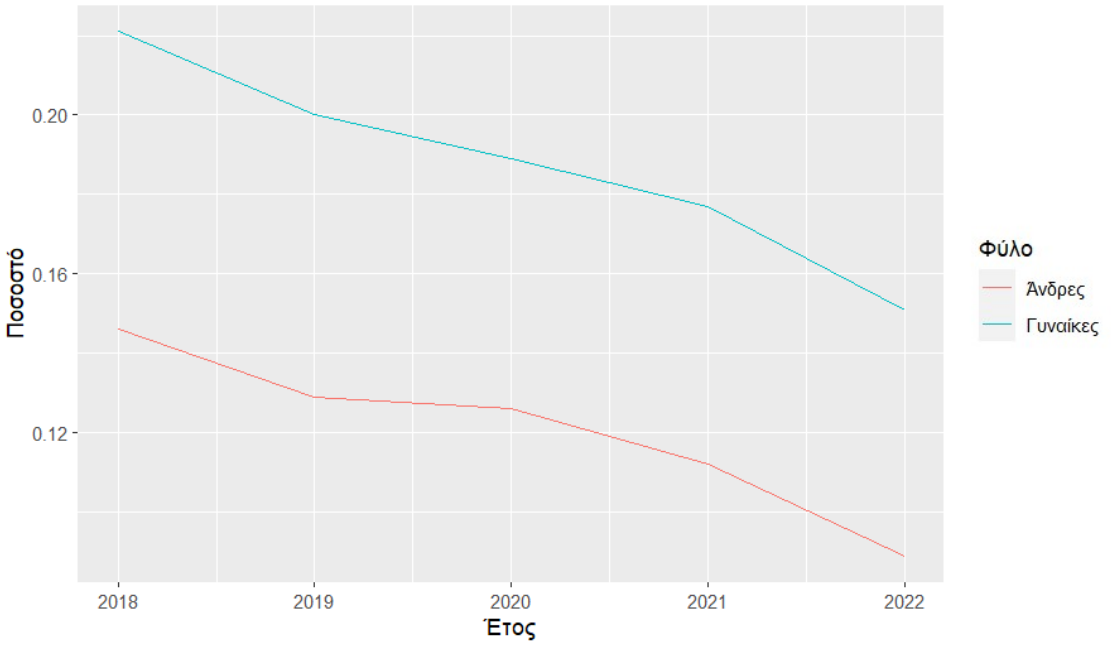
Σχήμα 2.19: Η εξέλιξη της ανεργίας από το 2018 έως το 2022

2.4.1 Ως προς το φύλο

Ελέγχοντας την ανεργία στους άνδρες παρατηρείται σταθερή μείωση από το 2018 έως το 2022. Το ίδιο παρατηρείται και στο ποσοστό της ανεργίας των γυναικών, όπως φαίνεται στον Πίνακα 2.41 και στο Σχήμα 2.20.

Έτος	Άνδρες	Γυναίκες
2018	14,6%	22,1%
2019	12,9%	20%
2020	12,6%	18,9%
2021	11,2%	17,7%
2022	8,9%	15,1%

Πίνακας 2.41: Ποσοστό ανεργίας ανά φύλο από το 2018 έως το 2022



Σχήμα 2.20: Η εξέλιξη της ανεργίας ανά φύλο από το 2018 έως το 2022

2.4.2 Ως προς την ηλικία

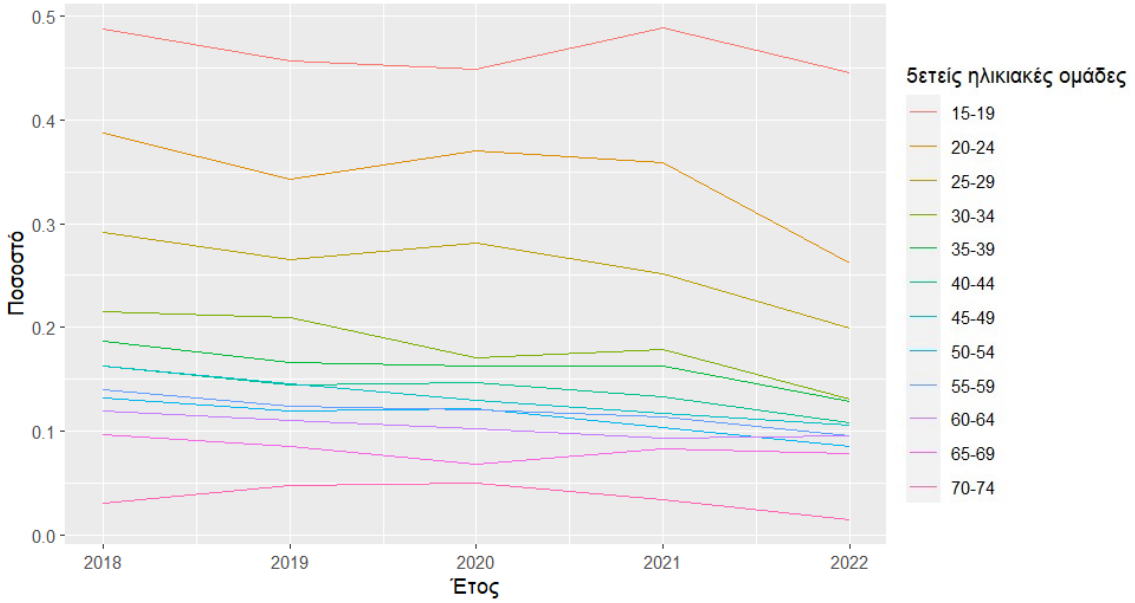
Όπως παρατηρούμε από τους πίνακες 2.42, 2.43 και από το Σχήμα 2.21, η ανεργία για τις 12 5ετείς ηλικιακές ομάδες παρουσιάζει μείωση από το 2018 έως το 2022. Αύξηση παρουσιάζει η ανεργία των ατόμων ηλικίας 15-19 και 30-34 για το έτος 2021, η οποία ακολουθείται από μείωση το επόμενο έτος (2022). Αύξηση παρουσιάζει και η ανεργία των ατόμων 70-74 ετών για τα έτη 2019 έως 2021, όπου το 2022 φτάνει το μικρότερο σημείο της.

Έτος	**15-19**	**20-24**	**25-29**	**30-34**	**35-39**	**40-44**
2018	48,7%	38,7%	29,1%	21,5%	18,7%	16,3%
2019	45,7%	34,3%	26,6%	21,0%	16,7%	14,4%
2020	44,9%	37,0%	28,1%	17,1%	16,3%	14,7%
2021	48,8%	35,9%	25,1%	17,8%	16,3%	13,3%
2022	44,5%	26,2%	19,9%	13,1%	12,9%	10,8%

Πίνακας 2.42: Ποσοστό ανεργίας ανά ηλικιακή ομάδα από το 2018 έως το 2022

Έτος	45-49	50-54	55-59	60-64	65-69	70-74
2018	16,3%	13,2%	14,0%	11,9%	9,7%	3,1%
2019	14,5%	12,0%	12,5%	11,0%	8,5%	4,8%
2020	13,0%	12,2%	12,1%	10,2%	6,8%	5,0%
2021	11,7%	10,4%	10,4%	9,4%	8,3%	3,5%
2022	10,6%	8,6%	9,5%	9,5%	7,9%	1,5%

Πίνακας 2.43: Ποσοστό ανεργίας ανά ηλικιακή ομάδα από το 2018 έως το 2022



Σχήμα 2.21: Η εξέλιξη της ανεργίας ανά ηλικιακή ομάδα από το 2018 έως το 2022

2.4.3 Ως προς το επίπεδο εκπαίδευσης

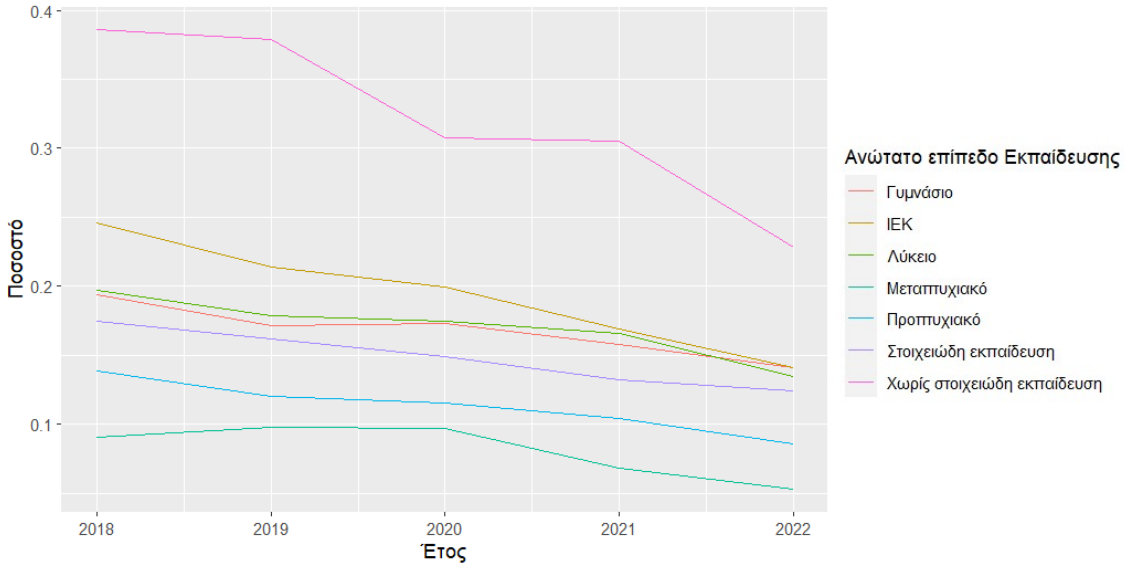
Όπως παρατηρείται στον Πίνακα 2.44 και 2.45, καθώς και στο Σχήμα 2.22, τα ποσοστά της ανεργίας για κάθε ανώτατο επίπεδο ολοκληρωμένης εκπαίδευσης σημειώνουν σημαντική μείωση από το 2018 έως το 2022.

Έτος	Χωρίς Στοιχειώδη Εκπαίδευση	Στοιχειώδη Εκπαίδευση	Γυμνάσιο
2018	38,6%	17,5%	19,4%
2019	37,9%	16,2%	17,1%
2020	30,7%	14,9%	17,3%
2021	30%	13,2%	15,8%
2022	22,8%	12,4%	14,1%

Πίνακας 2.44: Ποσοστό ανεργίας ανά ανώτατο επίπεδο εκπαίδευσης από το 2018 έως το 2022

Έτος	Λύκειο	ΙΕΚ	Προπτυχιακό	Μεταπτυχιακό
2018	19,7%	24,6%	13,8%	9,1%
2019	17,9%	21,4%	12,0%	9,7%
2020	17,5%	20%	11,5%	9,7%
2021	16,6%	16,9%	10,40%	6,8%
2022	13,4%	14,1%	8,6%	5,3%

Πίνακας 2.45: Ποσοστό ανεργίας ανά ανώτατο επίπεδο εκπαίδευσης από το 2018 έως το 2022



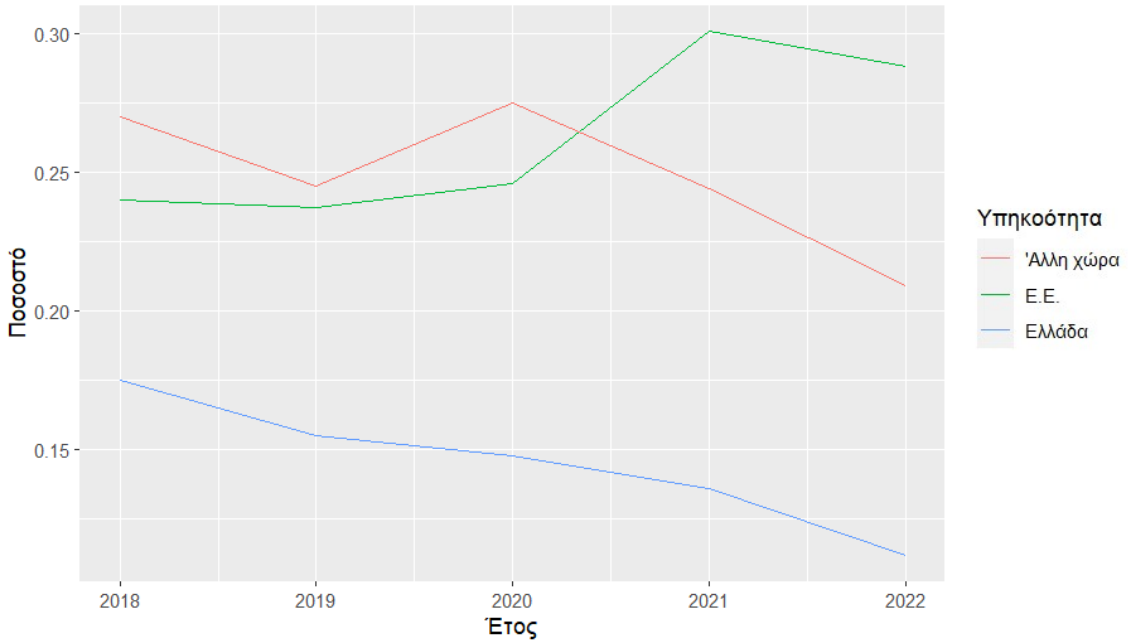
Σχήμα 2.22: Η εξέλιξη της ανεργίας ανά ανώτατο επίπεδο εκπαίδευσης από το 2018 έως το 2022

2.4.4 Ως προς την υπηκοότητα

Η ανεργία των ατόμων με Ελληνική υπηκόοτητα παρουσιάζει μείωση από το 2018 έως το 2022, όπως φαίνεται από τον Πίνακα 2.46 και Σχημά 2.23. Η ανεργία των ατόμων με υπηκόοτητα από χώρα της Ευρωπαϊκής Ένωσης παρουσιάζει μείωση το 2019 σε σχέση με το 2018, όμως έκτοτε αυξάνεται. Το 2022 σημειώνει μείωση συγκριτικά με το 2021, όμως παραμένει αυξημένο σε σχέση με το 2018. Τέλος, η ανεργία των ατόμων με υπηκοότητα από Άλλη Χώρα μειώνεται από το 2018 έως το 2022, με εξαίρεση το 2020 που παρουσιάζει αύξηση.

Έτος	Ελλάδα	Χώρα Ευρωπαϊκής Ένωσης	Άλλη χώρα
2018	17,5%	24%	27%
2019	15,5%	23,7%	24,5%
2020	14,8%	24,6%	27,5%
2021	13,6%	30,06%	24,4%
2022	11,2%	28,8%	20,9%

Πίνακας 2.46: Ποσοστό ανεργίας ανά υπηκοότητα από το 2018 έως το 2022



Σχήμα 2.23: Η εξέλιξη της ανεργίας ανά υπηκοότητα από το 2018 έως το 2022

2.4.5 Ως προς την αστικοποίηση

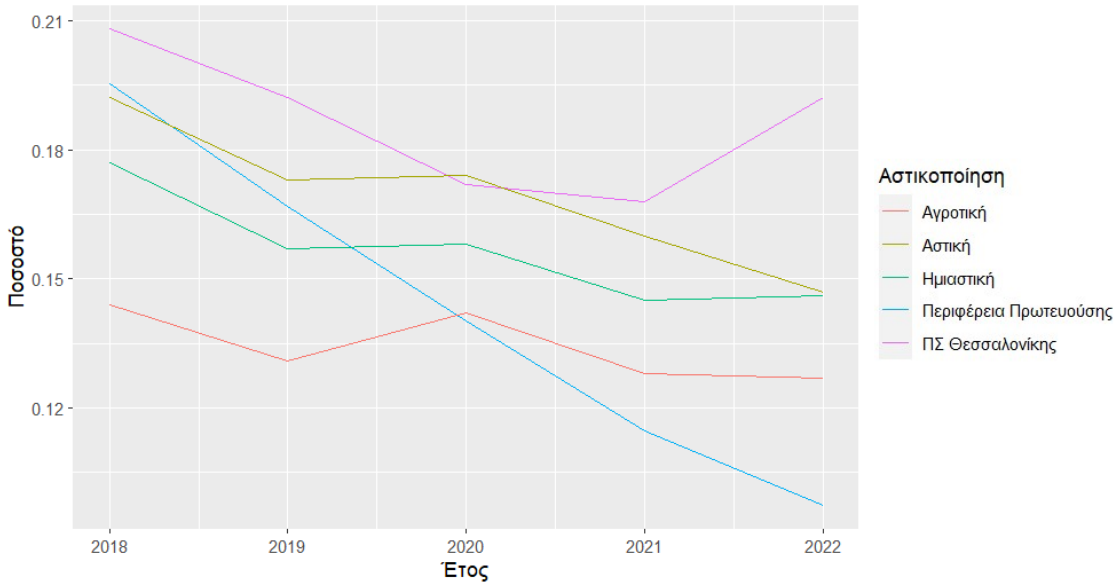
Όσον αφορά την ανεργία στους 5 βαθμούς αστικοποίησης, παρατηρείται από τον Πίνακα 2.47, 2.48 και το Σχήμα 2.24 ότι η Περιφέρεια Πρωτευούσης, οι Αστικές και οι Αγρότικές περιοχές παρουσιάζουν μείωση της ανεργίας από το 2018 έως το 2022. Στο Πολεοδομικό Συγκρότημα της Θεσσαλονίκης (ΠΣ Θεσσαλονίκης) το ποσοστό της ανεργίας μειώνεται από το 2018 έως το 2021, ενώ το 2022 αυξάνεται και βρίσκετα στο ίδιο ποσοστό με το 2019. Στις Ημιαστικές περιοχές παρουσιάζεται μείωση της ανεργίας από το 2018 έως το 2021, ενώ το 2022 αυξάνεται ελάχιστα.

Έτος	Περιφ. Πρωτευούσης	ΠΣ Θεσσαλονίκης
2018	20,2%	20,8%
2019	17,8%	19,2%
2020	13,8%	17,2%
2021	11,8%	16,8%
2022	10,4%	19,2%

Πίνακας 2.47: Ποσοστό ανεργίας ανά αστικοποίηση από το 2018 έως το 2022

Έτος	Αστική	Ημιαστική	Αγρότική
2018	19,2%	17,7%	14,4%
2019	17,3%	15,7%	13,09%
2020	17,4%	15,8%	14,2%
2021	16%	14,5%	12,8%
2022	14,7%	14,6%	12,7%

Πίνακας 2.48: Ποσοστό ανεργίας ανά αστικοποίηση από το 2018 έως το 2022



Σχήμα 2.24: Η εξέλιξη της ανεργίας ανά βαθμό αστικοποίησης από το 2018 έως το 2022

2.4.6 Ως προς τις περιφέρειες

Εξετάζοντας την ανεργία στις 13 Περιφέρειες της Ελλάδος παρατηρείται, όπως φαίνεται στους Πίνακες 2.49, 2.50, 2.51, 2.52, 2.53 και στο Σχήμα 2.25, τα ποσοστά ανεργίας για τις Περιφέρεις της ανάτολικής Μακεδονίας και Θράκης, της Κεντρικής Μακεδονίας, της Δυτικής Μακεδονίας, της Ηπείρου, της Θεσσαλίας, της Δυτικής Ελλάδας, της Στερεάς Ελλάδας, της Αττικής και του Βόρειου Αιγαίου μειώνονται από το 2018 έως το 2022.

Εξαίρεση αποτελεί η Περιφέρεια των Ιόνιων Νησιών, όπου το ποσοστό της ανεργίας αυξήθηκε το 2020 (σε σύγκριση με τις προηγούμενες χρονιές, 2018-2019) και το 2021 και 2022 μειώθηκε, με το 2022 να σημειώνει το χαμηλότερο ποσοστό του. Επιπλέον, στην Περιφέρεια του Νοτίου Αιγαίου παρατηρείται αύξηση για το έτος 2020, όπως και στην Περιφέρεια της Κρήτης, με τα ποσοστά να μειώνονται εκ νέου το 2021 και 2022.

Έτος	**Α.Μακεδονία και Θράκη**	**Κεντ. Μακεδονία**
2018	14,8%	20,2%
2019	15,4%	19,3%
2020	15,5%	19,3%
2021	14,3%	15,0%
2022	10,9%	13,7%

Πίνακας 2.49: Ποσοστό ανεργίας ανά περιφέρεια από το 2018 έως το 2022

Έτος	**Δυτ.Μακεδονία**	**Ήπειρος**	**Θεσσαλία**
2018	25,1%	18,3%	16,2%
2019	22,2%	16,1%	15,4%
2020	18,3%	18,0%	13,6%
2021	15,9%	14,9%	13,5%
2022	14,2%	13,3%	13,3%

Πίνακας 2.50: Ποσοστό ανεργίας ανά περιφέρεια από το 2018 έως το 2022

Έτος	**Ιόνια Νησιά**	**Δυτική Ελλάδα**	**Στερεά Ελλάδα**
2018	13,4%	21,7%	17,2%
2019	11,2%	21,1%	15,1%
2020	12,9%	17,5%	15,8%
2021	11,9%	16,5%	13,3%
2022	5,7%	11,2%	12,0%

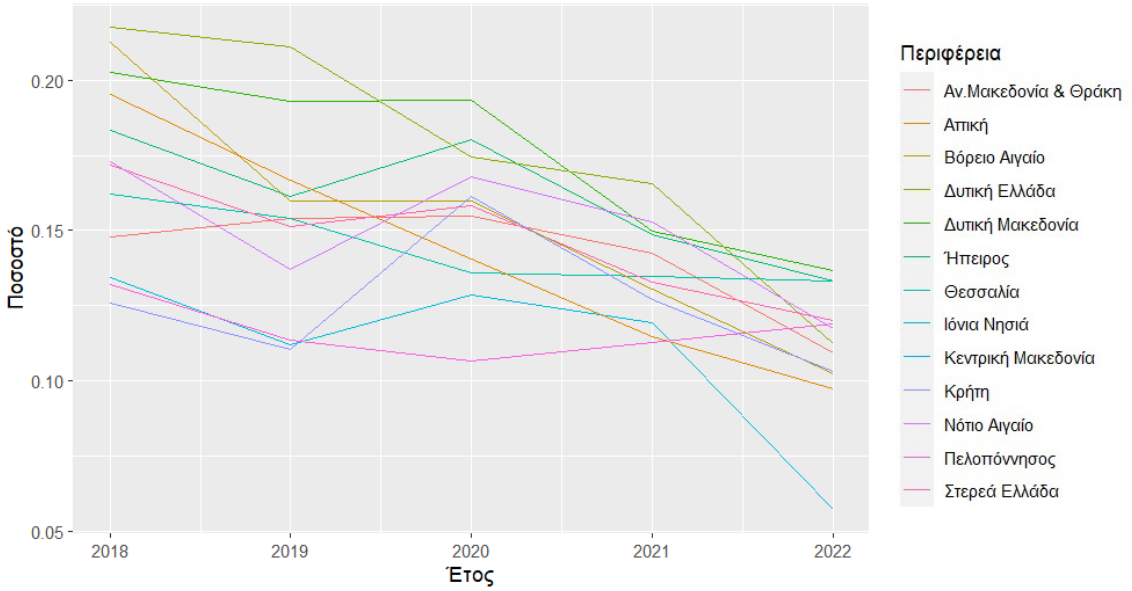
Πίνακας 2.51: Ποσοστό ανεργίας ανά περιφέρεια από το 2018 έως το 2022

Έτος	**Αττική**	**Πελοπόννησος**	**Βόρειο Αιγαίο**
2018	19,5%	13,2%	21,3%
2019	16,7%	11,3%	16,0%
2020	14,0%	10,6%	16,0%
2021	11,5%	11,3%	13,1%
2022	9,7%	11,9%	10,2%

Πίνακας 2.52: Ποσοστό ανεργίας ανά περιφέρεια από το 2018 έως το 2022

Έτος	Νότιο Αιγαίο	Κρήτη
2018	17,3%	12,6%
2019	13,7%	11,1%
2020	16,8%	16,1%
2021	15,3%	12,7%
2022	11,7%	10,3%

Πίνακας 2.53: Ποσοστό ανεργίας ανά περιφέρεια από το 2018 έως το 2022



Σχήμα 2.25: Η εξέλιξη της ανεργίας ανά περιφέρεια από το 2018 έως το 2022

2.4.7 Συμπεράσματα

Το ποσοστό της ανεργίας για την Ελλάδα το 2022 (11,7%) είναι από τα υψηλότερα της Ευρώπης, όμως παρατηρείται σταδιακή μείωση της ανεργίας τα τελευταία έτη για τον Ελλαδικό χώρο. Ανάμεσα στα δύο φύλα παρατηρείται ότι οι άνδρες έχουν περισσότερες ευκαιρίες απασχόλησης, ενώ οι περισσότερες γυναίκες ανήκουν στα άτομα εκτός εργατικού δυναμικού. Όπως είδαμε τα άτομα που αποτελούν τα εκτός εργατικού δυναμικού είναι μεγαλύτερης ηλικίας, με μέσο όρο ηλικίας τα 63,2 έτη, όπου το μεγαλύτερο μέρος τους έχει λάβει την στοιχειώδη εκπαίδευση και ζεί σε αγρότικές περιοχές, κυρίως στην Περιφέρεια της Κρήτης και της Πελοποννήσου.

Στο σύνολο των εργαζομένων οι περισσότεροι εργάζονται υπό το καθεστώς πλήρους απασχόλησης και με σύμβαση αορίστου χρόνου, χωρίς ιδιαίτερες διαφορές ανάμεσα στα δύο φύλα, ούτε στις 3 υπηκοότητες. Όσον αφορά την κύρια θέση στο σύνολο υπερτερούν οι θέσεις των μισθωτών,

με τους περισσότερους να βρίσκονται σε αγρότικές περιοχές και να είναι απόφοιτοι λυκείου. Ενώ οι περισσότεροι αυτοαπασχολούμενοι είναι σε αστικές περιοχές και έχουν ολοκληρώσει το λύκειο, και τέλος οι βοηθοί σε οικογενειακές επιχειρήσεις είναι κυρίως σε αγρότικές περιοχές με ανώτατοι βαθμίδα εκπαίδευσης την τριτοβάθμια. Η Ελλάδα κατέχει μια από τις υψηλότερες θέσεις στις περισσότερες μέσες εβδομαδιαίες ώρες εργασίας, με τους άνδρες να εργάζονται περισσότερο, όπως επίσης και τα άτομα που διαμένουν σε αστικές περιοχές. Ο κλάδος με την μεγαλύτερη απορρόφηση αφορά το χονδρικό και λιανικό εμπόριο, και συγκεκριμένα την παροχή υπηρεσιών και τις πωλήσεις. Οι περισσότεροι άνδρες απορροφώνται σε τέτοιου είδους θέσεις, ενώ οι γυναίκες εργάζονται κυρίως σαν επαγγελματίες και σχεδόν καθόλου σε χειρωνακτικά επαγγέλματα.

Οι άνεργοι στο σύνολο τους είναι απόφοιτοι λυκείου, διαμένουν σε αστικές περιοχές, οι οποίοι έχουν να εργαστούν πάνω από ένα χρόνο διότι η τελευταία εργασία τους ήταν περιορισμένης διάρκειας, ανεξάρτητα από το φύλο και αναζητούν εργασία κατα κύριο λόγο πάνω από 12 μήνες. Το μεγαλύτερο ποσοστό αυτών λαμβάνει επίδομα και είναι εγγεγραμένο σε πρόγραμμα αναζήτησης εργασίας.

Κεφάλαιο 3

Λογιστική Παλινδρόμηση

3.1 Θεωρητικό Υπόβαθρο Λογιστικής Παλινδρόμησης

Το μοντέλο της λογιστικής παλινδρόμησης αποτελεί μια ειδική περίπτωση των γενικευμένων γραμμικών μοντέλων. Η λογιστική παλινδρόμηση είναι η κατάλληλη μέθοδος όταν η μεταβλητή απόκρισης του μοντέλου αποτελείται από δύο τιμές, συνήθως με την κωδικοποίηση 0 ή 1. Οι δυό αυτές τιμές αντιστοιχούν σε δύο διαφορετικά ενδεχόμενα (σύνηθως το ένα αποτελεί την επιτυχία και το άλλο την αποτυχία). Γίνεται αντιληπτό πως η μεταβλητή απόκρισης Y είναι τυχαία μεταβλητή (τ.μ.) της κατανομής *Bernoulli*, η οποία αποτελεί διακριτή πιθανότητα με δύο πιθανές τιμές, άρα $Y \sim B(p)$ με $p \in (0,1)$. Η p εκφράζει την πιθανότητα επιτυχίας όταν η μεταβλητή απόκρισης, Y, λάβει την τιμή 1 και αφού τα δύο ενδεχόμενα είναι συμπληρωματικά η $1\text{-}p$ εκφράζει την πιθανότητα αποτυχίας όταν η $Y=0$. Επομένως, αφού η $Y \sim B(p)$ θα ισχύει ότι η μέση τιμή της Y θα ισούται με $E(Y)=p$ και η διασπορά της με $V(Y)=p(1\text{-}p)$. Η μέση τιμή της διακριτής τυχαίας μεταβλητής Y προκύπτει ως:

$$E(Y) = \sum_{i=1}^{n} y_i \cdot P(Y = y_i),$$

και αφού η μεταβλητή y_i παίρνει δύο πιθανές τιμές 0 ή 1:

$$E(Y) = 0 \cdot (1 - p_i) + 1 \cdot p_i = p_i.$$

Η διασπορά της κατανομής Bernoulli προκύπτει ως εξής:

$$V(Y) = E(y_i - E(y_i))^2 = (1 - p_1)^2 p_i + (0 - p_i)^2 (1 - p_i) = p_i(1 - p_i).$$

Από την παραπάνω εξίσωση αντιλαμβανόμαστε ότι η διασπορά είναι μια συνάρτηση που εξαρτάται από την μέση τιμή, αφού

$$V(Y) = E(y_i)(1 - E(y_i)).$$

Αν επεκτείνουμε την πραγματοποίηση των γεγονότων σε μια σειρά από n δοκιμές, δηλαδή η Y αποτελεί των αριθμό επιτυχιών σε n δοκιμές. Αν επιπλέον υποθέσουμε ότι η πιθανότητα επιτυχίας p είναι ίδια για κάθε μια από τις n δοκιμές και οι δοκιμές είναι ανεξάρτητες μεταξύ τους, τότε ισχύει η Διωνυμική κατανομή με $Y \sim b(n,p)$ και η μέση τιμή της ισούται με $E(Y)=np$ και η διασπορά της με $V(Y)=np(1-p)$. Στην περίπτωση της παρούσας εργασίας πρόκειται για δυαδικά δεδομένα, επομένως το $n=1$ (μια δοκιμή).

Ο στόχος του μοντέλου της λογιστικής παλινδρόμησης θα είναι η εξάρτηση μεταξύ της μεταβλητής απόκρισης, Y και των επεξηγηματικών μεταβλητών τις οποίες συμβολίζουμε με το διάνυσμα $\boldsymbol{X}$. Η εξάρτηση θα εισαχθεί μέσω της εξάρτησης της πιθανότητας επιτυχίας από το $\boldsymbol{X}$, εν ολίγοις να προβλέψουμε την πιθανότητα επιτυχίας της Y.

Αν υποθέσουμε ότι η μορφή του μοντέλου είναι γραμμική, θα είχαμε ένα μοντέλο της μορφής $y_i = x_i'b + \varepsilon_i$. Επομένως, η μέση τιμή των παρατηρήσεων y_i θα ήταν $E(y_i) = x_i'b = p_i$, με την μέση τιμή των σφαλμάτων ίση με 0 ($E(\varepsilon_i) = 0$). Γνωρίζουμε όμως ότι οι παρατηρήσεις y_i μπορούν να πάρουν δύο πιθανές τιμές 0 ή 1, και επομένως τα σφάλματα θα είχαν τις εξής τιμές: Για $y_i = 1$ Θα ήταν $\varepsilon_i = 1 - x_i'b$ και για $y_i = 0$ το σφάλμα θα ήταν $\varepsilon_i = 0 - x_i'b$. Με αυτό τον τρόπο δεν ισχύει η υπόθεση της κανονικής κατανομής των σφαλμάτων, μια βασική προυπόθεση του γραμμικού μοντέλου.

Ένας ακόμη πολύ σημαντικός περιορισμός που προκύπτει από αυτή την μορφή του μοντέλου είναι ότι η μέση τιμή ισούται με την πιθανότητα, $E(y_i) = p_i$. Αυτή η σχέση περιορίζει την μέση τιμή να λαμβάνει τιμές στο $[0, 1]$, όμως γνωρίζουμε πως οι τιμές $E(y_i) = x_i'b$ λαμβάνει τιμές σε όλο το $\mathbb{R}$.

Επομένως, για να ξεπεράσουμε τον παραπάνω περιορισμό ψάχνουμε μια συνάρτηση σύνδεσης, g, τέτοια ώστε:

$$g = \mathbb{E}[Y|\mathbf{X} = x] = \mathbb{P}[Y = 1|\mathbf{X} = x] = p = \frac{exp(a + b_1x_1 + \ldots + b_px_p)}{1 + exp(a + b_1x_1 + \ldots + b_px_p)}. \tag{3.1}$$

Η συνάρτηση 3.1 δεν είναι γραμμική, αλλά αυτό μπορεί εύκολα να επιτευχθεί.

Άρα, η συμπληρωματική πιθανότητα

$$\mathbb{P}[Y = 0|\mathbf{X} = x] = 1 - p = 1 - \frac{exp(a + b_1x_1 + ... + b_px_p)}{1 + exp(a + b_1x_1 + ... + b_px_p)}. \quad (3.2)$$

Παίρνοντας τον λόγο των συμπληρωματικών πιθανοτήτων, δηλαδή των εξισώσεων (3.1) και (3.2):

$$\frac{\mathbb{P}[Y = 1|\mathbf{X} = x]}{\mathbb{P}[Y = 0|\mathbf{X} = x]} = exp(a + b_1x_1 + ... + b_px_p). \quad (3.3)$$

Ο παραπάνω λόγος ονομάζεται σχετική πιθανότητα (odds).

Στην συνέχεια, εφαρμόζοντας λογάριθμο στην σχέση (3.3):

$$log(\frac{\mathbb{P}[Y = 1|\mathbf{X} = x]}{\mathbb{P}[Y = 0|\mathbf{X} = x]}) = a + b_1x_1 + ... + b_px_p. \quad (3.4)$$

Άρα, καταλήγουμε στο μοντέλο:

$$log(\frac{p}{1 - p}) = a + b_1x_1 + ... + b_px_p. \quad (3.5)$$

Η συνάρτηση (3.4), και συνεπώς η (3.5), αποτελεί την συνάρτηση *logit* η οποία αποτελεί τη συνάρτηση σύνδεσης. Το μοντέλο μετασχηματίζει την εξίσωση της δεσμευμένης πιθανότητας, στον λόγο των συμπληρωματικών πιθανοτήτων (odds) και τέλος λογαριθμίζει την εξίσωση, με αυτό τον τρόπο επιτυγχάνεται το κατώτερο και το ανώτερο όριο της πιθανότητας να αφαιρούνται και η μέση τιμή να λαμβάνει πλέον τιμές σε όλο το $\mathbb{R}$.

Αν η προσέγγιση της πιθανότητας γινόταν μέσω της γραμμικής παλινδρόμησης αυτό θα οδηγούσε οι προβλεπόμενες τιμές θα ήταν εκτός του διαστήματος $(0,1)$, καθώς η γραμμική συνάρτηση των επεξηγηματικών μεταβλητών θα έπαιρνε τιμές σε όλο το $\mathbb{R}$.

Από την εξίσωση 3.5 αντιλαμβανόμαστε ότι η πιθανότητα, p, θα ισούται με:

$$p = \frac{exp(x_i'b)}{1 + exp(x_i'b)}.$$

3.1.1 Προϋποθέσεις Λογιστικού Μοντέλου

Σε αντιδιαστολή με το απλό γραμμικό μοντέλο παλινδρόμησης, για την εφαρμογή του λογιστικού μοντέλου απαιτούνται λιγότερες προϋποθέσεις. Αρχικά, εκτός της προφανής προϋπόθεσης ότι η μεταβλητή απόκρισης, Y, πρέπει να είναι δίτιμη (με τιμές της μορφής 0/1), δεν προϋποθέτει την ύπαρξη της κανονικής κατανομής (αφού πρόκειται για διακριτή κατανομή). Επιπλεόν, λόγω της δίτιμης φύσης της μεταβλητής παραβιάζεται η ομοσκεδαστικότητα (ότι όλες οι παρατηρήσεις έχουν την ίδια διασπορά), αφού κάθε παρατήρηση έχει διαφορετική διασπορά, $V(y_i) = p_i(1 - p_i)$. Οι βασικές προϋποθέσεις του μοντέλου είναι η γραμμικότητα και η ανεξαρτησία. Στην γραμμικότητα το ζητούμενο είναι η γραμμική σχέση μεταξύ της λογαριθμικής σχετικής πιθανότητας της μεταβλητής απόκρισης, Y και των ποσότικών επεξηγηματικών μεταβλητών. Η ανεξαρτησία επιτυγχάνεται όταν οι παρατηρήσεις είναι ανεξάρτητες τόσο μεταξύ τους, όσο και ανεξάρτητες από επαναλμβανόμενες μετρήσεις. Κάθε μονάδα του δείγματος θα πρέπει να μετράται μόνο μια φορά, διαφορετικά δεν θα λαμβάνεται υπόψιν στο μοντέλο. Ένα επιπλεόν ζήτημα που μπορεί να προκύψει και χρήζει αντιμετώπιση είναι η πολυσυγγραμμικότητα. Το φαινόμενο αυτό εμφανίζεται όταν υπάρχει ισχυρή συσχέτιση μεταξύ δύο ή περισσοτέρων επεξηγηματικών μεταβλητών.

3.1.2 Μέθοδος και Εκτίμηση Παραμέτρων

Στο μοντέλο της λογιστικής παλινδρόμησης η προσαρμογή του μοντέλου στα δεδομένα γίνεται με την μέθοδο της μεγίστης πιθανοφάνειας. Μέσω της μεθόδου μεγίστης πιθανοφάνειας εκτιμάται η διανυσματική παράμετρος $\boldsymbol{\Theta}$ που αποτελείται από τους συντελεστές του μοντέλου (αν θεωρήσουμε πως το μοντέλο περιέχει p επεξηγηματικές μεταβλητές) $\mathbf{X} = (X_1, \ldots, X_p)'$), $\boldsymbol{\Theta} = (a, b_1, \ldots, b_p)'$. Έστω ότι διαθέτουμε το τυχαίο δείγμα,

$$(Y_1, X_{11}, \ldots, X_{1p})', \ldots, (Y_p, X_{p1}, \ldots, X_{pp})'$$

με τιμές,

$$(y_1, x_{11}, \ldots, x_{1p})', \ldots, (y_p, x_{p1}, \ldots, x_{pp})'$$

και επιπλέον το διάνυσμα των επεξηγηματικών μεταβλητών,

$$\mathbf{X_i} = (X_{i1}, \ldots, X_{ip})'$$

με γνωστή τιμή $\mathbf{x_i} = (x_{i1}, \ldots, x_{ip})'$ όπου $i = 1, \ldots, n$.

Θεωρούμε πως κάθε παρατήρηση ακολουθεί την κατανομή Bernoulli, οπότε παίρνουμε τη συνάρτηση μάζας πιθανότητας (σ.μ.π) της κατανομής,

$$f(x;\boldsymbol{\Theta}) = p_i^{y_i}(1-p_i)^{1-y_i},$$

άρα η απο κοινού σ.μ.π. του τυχαίου δείγματος, αφού θεωρούμε ότι κάθε παρατήρηση είναι ανεξάρτητη, θα είναι:

$$L(\boldsymbol{\Theta}) = \prod_{i=1}^{n} p_i^{y_i}(1-p_i)^{1-y_i}.$$

Λογαριθμίζοντας την παραπάνω συνάρτηση,

$$log(L(\boldsymbol{\Theta})) = \sum_{i=1}^{n}\Big(y_i log(p_i) + (1-y_i)log(1-p_i)\Big)$$

$$= \sum_{i=1}^{n}\Big(y_i log(p_i) + log(1-p_i) - y_i log(1-p_i)\Big)$$

$$= \sum_{i=1}^{n}\Big(y_i log\Big(\frac{p_i}{1-p_i}\Big) + log(1-p_i)\Big)$$

$$== \sum_{i=1}^{n}\Big(y_i log\Big(exp(x'\boldsymbol{\Theta})\Big) + log(1-p_i)\Big)$$

$$= \sum_{i=1}^{n}\Big(y_i x_i'\boldsymbol{\Theta} + log(1-p_i)\Big)$$

$$= \sum_{i=1}^{n}\Big(y_i x_i'\boldsymbol{\Theta} + log(1 - \frac{exp(x_i'\boldsymbol{\Theta})}{1+exp(x_i'\boldsymbol{\Theta})})\Big)$$

$$= \sum_{i=1}^{n}\Big(y_i x_i'\boldsymbol{\Theta} - log(1+exp(x_i'\boldsymbol{\Theta}))\Big). \qquad (3.6)$$

Παραγωγίζοντας την λογαριθμική συνάρτηση πιθανοφάνειας ως προς τις παραμέτρους του μοντέλου και εξισώνοντας με το 0, προκύπτει :

$$\frac{\partial log(\boldsymbol{\Theta})}{\partial \beta_j} = \sum_{i=1}^{n}(y_i x_{ij}) - \sum_{i=1}^{n} x_{ij} exp(x_i'\partial)(1+exp(x_i'\partial)^{-1}),$$

$$= \sum_{i=1}^{n}\Big(y_i - exp(x_i'\partial)(1+exp(x_i'\partial)^{-1}\Big)x_{ij}$$

$$= \sum_{i=1}^{n} \left(y_i - p_i\right)x_{ij}$$

όπου $j = 0, \ldots, p$. Άρα, προκύπτουν οι εξισώσεις:

$$\sum_{i=1}^{n} \left(y_i - p_i\right)x_{ij} = 0. \tag{3.7}$$

Λύνοντας τις $(p + 1)$ (3.7) μη γραμμικές εξισώσεις με επαναληπτικές μεθόδους οδηγούμαστε στις τιμές των εκτιμητριών του $\boldsymbol{\Theta}$.

3.1.3 Ερμηνεία Συντελεστών

Η ερμηνεία των συντελεστών ενός μοντέλου λογιστικής παλινδρόμησης είναι εφικτή και σχετικά παρόμοια με εκείνη του πολλαπλού γραμμικού μοντέλου παλινδρόμησης.

Έστω το μοντέλο

$$log(\frac{p_i}{1 - p_i}) = a + b_1x_1 + \ldots + b_px_p.$$

Παρατηρείται ότι η επίδραση των επεξηγηματικών μεταβλητών, b_j, υπολογίζεται ως προς των λογάριθμο του λόγου των συμπληρωματικών πιθανοτήτων ($log(odds)$).

Η σταθερά a εκφράζει την τιμή του $log(odds)$ για την εμφάνιση της τιμής "επιτυχίας" της μεταβλητής απόκρισης Y, όταν οι συνεχείς επεξηγηματικές μεταβλητές έχουν μηδενικές τιμές και οι κατηγορικές μεταβλητές ισούται με την κατηγορία αναφοράς.

Οι συντελεστές b_j με $j = 1, \ldots, p$ στην περίπτωση των ποσότικών επεξηγηματικών μεταβλητών εκφράζουν την μεταβολή του $log(odds)$ για κάθε μονάδα που μεταβάλλεται η αντίστοιχη επεξηγηματική μεταβλητή, ενώ οι υπόλοιπες επεξηγηματικές παραμένουν σταθερές. Άρα, ο συντελεστής b_j ισούται με $log(odds(x_j + 1) - log(odds(x_j)) = b_j$.

Στην περίπτωση των κατηγορικών επεξηγηματικών μεταβλητών, οι συντελεστές b_j εκφράζουν την τιμή του $log(odds)$ των υπόλοιπων κατηγοριών της μεταβλητής ως προς την κατηγορία αναφοράς, ενώ οι υπόλοιπες μεταβλητές παραμένουν σταθερές. Όταν μια κατηγορική μεταβλητή έχει πάνω από δύο κλάσεις, δημιουργούμε εικονικές μεταβλητές και η ανάλυση γίνεται με αντίστοιχο τρόπο. Στην R η δημιουργία εικονικών μεταβλητών (dummy variables) γίνεται αυτόματα κατα την προσαρμογή

του μοντέλου. Ο αριθμός των εικονικών μεταβλητών είναι πάντα $m-1$, όπου m είναι ο αριθμός των κλάσεων της κατηγορικής μεταβλητής.

Η συνάρτηση *logit* μπορεί να είναι περίπλοκη, για αυτό το λόγο μπορούμε να ξανά γράψουμε το μοντέλο ώστε να ερμηνεύεται ως προς το *odds*, τον λόγο των συμπληρωματικών πιθανοτήτων. Αυτό επιτυγχάνεται με την εκθετική συνάρτηση ως εξής:

$$\frac{p}{1-p} = exp(x'\boldsymbol{\Theta}) = exp(a + b_1x_1 + ... + b_px_p).$$

Επομένως, η ποσότητα $exp(b_j)$ είναι ο παράγοντας με τον οποίο πολλαπλασιάζεται ο όρος *odds*, όταν η αντίστοιχη ανεξάρτηση μεταβλητή, X μεταβληθεί κατα μια μονάδα, ενώ οι υπόλοιπες ανεξάρτητες μεταβλητές παραμένουν σταθερές πάντα. Προφανώς, όταν ο $b_j > 0$ ο εκθετικός, $exp(b_j)$, είναι μεγαλύτερος της μονάδας. Επομένως, και ο λόγος των συμπληρωματικών πιθανοτήτων, $\frac{p}{1-p}$, αυξάνεται με την αύξηση της X. Ομοίως, αν ο $b_j < 0$ τότε ο εκθετικός όρος, $exp(b_j) < 1$, είναι μικρότερος της μονάδας και άρα ο λόγος των συμπληρωματικών πιθανοτήτων, $\frac{p}{1-p}$, μειώνεται με την αύξηση της X.

3.1.4 Κριτήρια Καλής Προσαρμογής

Ένα πολύ σημαντικό ζήτημα στην ανάλυση ενός μοντέλου είναι η εκτίμηση της καταλληλότητας του, δηλαδή κατα πόσο το μοντέλο περιγράφει επαρκώς τα δεδομένα. Όσον αφορά τα μοντέλα της λογιστικής παλινδρόμησης υπάρχουν διάφορα κριτήρια και συντελεστές που χρησιμοποιούνται για την βέλτιστη επιλογή του μοντέλο, καθώς και την καλή προσαρμογή. Από την σχέση (3.7) προέκυψαν οι εκτιμητές για το $\boldsymbol{\Theta}$, το διάνυσμα των οποίων συμβολίζουμε με $\hat{\boldsymbol{\Theta}}$. Οπότε αν χρησιμοποιήσουμε τους εκτιμητές, $\hat{\boldsymbol{\Theta}}$, στην θέση του $\boldsymbol{\Theta}$ στην λογαριθμική συνάρτηση πιθανοφάνειας, (3.6) προκύπτει η μέγιστη τιμή της συνάρτησης. Η μέγιστη τιμή θα ισούται με:

$$LL(\hat{\boldsymbol{\Theta}}) = \sum_{i=1}^{n} \Big(y_i x_i' \hat{\boldsymbol{\Theta}} - log(1 + exp(x_i' \hat{\boldsymbol{\Theta}})) \Big). \tag{3.8}$$

Κάνοντας χρήση της σχέσης (3.8), της μέγιστης τιμής $LL(\hat{\boldsymbol{\Theta}})$, αναπτύσσονται διάφορα κριτήρια για ένα λογιστικό μοντέλο. Αρχικά, η ίδια η $LL(\hat{\boldsymbol{\Theta}})$ μπορεί να χρησιμοποιηθεί σαν κριτήριο. Αν η μέγιστη τιμή της είναι κοντά στο μηδέν, τότε αυτό αποτελεί ένδειξη καλής προσαρμογής.

Αυτό συμβαίνει διότι η $LL(\hat{\boldsymbol{\Theta}})$ έχει πάντα αρνητική τιμή. Ο λόγος που η $LL(\hat{\boldsymbol{\Theta}})$ έχει πάντα αρνητική τιμή είναι ο εξής:

Ο όρος, $y_i x_i' \hat{\boldsymbol{\theta}}$ για $y_i = 0$ είναι προφανώς 0, ενώ για $y_i = 1$ είναι ίσος με $x_i' \hat{\boldsymbol{\theta}}$. Ο τελευταίος όρος είναι είτε θετικός, είτε αρνητικός ανάλογα με την πρόβλεψη για το $\hat{\boldsymbol{\theta}}$. Ο δεύτερος όρος της (3.8) είναι μεγαλύτερος από τον πρώτο, διότι η ποσότητα $exp(x_i' \hat{\boldsymbol{\theta}})$ είναι μεγαλύτερη από την $x_i' \hat{\boldsymbol{\theta}}$ και η ποσότητα $1 + exp(x_i' \hat{\boldsymbol{\theta}})$ πάντα μεγαλύτερη της μονάδας. Επομένως, η λογαριθμική ποσότητα $log(1 + exp(x_i' \hat{\boldsymbol{\theta}}))$ είναι μεγαλύτερη της $x_i' \hat{\boldsymbol{\theta}}$. Άρα, το άθροισμα της 3.8 είναι αρνητικό σε κάθε περίπτωση.

Ένα άλλο μέτρο που κάνει χρήση της $LL(\hat{\boldsymbol{\Theta}})$ είναι το μέτρο απόκλισης (Deviance), το οποίο ορίζεται ως:

$$D = -2 \times LL(\hat{\boldsymbol{\Theta}}).$$

Επομένως, το μέτρο της D είναι πάντα μη αρνητικό και για τιμές κοντά στο μηδέν έχουμε ένδειξη καλής προσαρμογής. Ακολουθούν τα κριτήρια AIC και BIC. Τα κριτήρια για ένα μοντέλο με p άγνωστες παραμέτρους και αριθμό παρατηρήσεων n, και κάνοντας πάλι χρήση της $LL(\hat{\boldsymbol{\Theta}})$ έχουν την μορφή:

$$AIC = -2 \times LL(\hat{\boldsymbol{\Theta}}) + 2p,$$

$$BIC = -2 \times LL(\hat{\boldsymbol{\Theta}}) + plog(n).$$

Τόσο με το κριτήριο AIC, όσο και με το BIC επιλέγεται το μοντέλο με την μικρότερη τιμή. Από τις σχέσεις των δύο κριτηρίων παρατηρούμε πως λαμβάνουν υπόψιν τους την προσαρμογή, αλλά και την διάσταση του μοντέλου.

Τέλος, υπάρχουν κριτήρια που βασίζονται στον συντελεστή προσδιορισμού R^2, τα οποία ονομάζονται ψευδο-R^2. Υπάρχουν τρείς τέτοιοι συντελεστές, οι οποίοι κάνουν χρήση της ελεγχοσυνάρτησης Deviance, D, και επομένως της

$$D = -2 \times LL(\hat{\boldsymbol{\Theta}}).$$

Όπως θα παρατηρήσουμε οι δείκτες χρησιμοποιούν εμφωλευμένα μοντέλα, έστω M_o και M_k. Το μοντέλο M_o αποτελείται μόνο από τον σταθερό όρο, ενώ το μοντέλο M_k περιέχει και p επεξηγηματικές μεταβλητές.

Αρχικά, ο δείκτης των *Homer και Lemeshow (1989)* ορίζεται ως:

$$R_L^2 = \frac{D(M_o) - D(M_k)}{D_o}.$$

Ο δείκτης των *Cox και Snell (1989)* ορίζεται ως:

$$R_C S^2 = 1 - exp(\frac{D(M_o) - D(M_k)}{n}).$$

Και τέλος, ο βελτιωμένος δείκτης του *Nagelkerke (1991)* ορίζεται ως:

$$R_N^2 = \frac{R_C S^2}{1 - exp\Big(- \frac{D(M_O)}{n}\Big)}.$$

Γνωρίζουμε πως ο συντελεστής προσδιορισμού, $R^2 \in [0, 1]$, το ίδιο ισχύει και στους παραπάνω συντελεστές. Άρα, αν η τιμή του συντελεστή είναι κοντά στο ένα υποδηλώνει ότι το μοντέλο M_k προσαρμόζεται καλύτερα σε σχέση με το μοντέλο M_o. Επομένως, για τιμές κοντά στο μηδέν σημαίνει πως η προσθήκη επεξηγηματικών μεταβλητών δεν συνεισφέρει στην καλύτερη προσαρμογή. Για την σύγκριση εμφωλευμένων μοντέλων, M_o και M_k, μπορούν να χρησιμοποιηθούν και τα κριτήρια πληροφορίας, AIC και BIC , καθώς και η ελεγχοσυνάρτηση Deviance. Συγκεκριμένα ελέγχουμε την διαφορά των ελεγχοσυναρτήσεων Deviance των δύο εμφωλευμένων μοντέλων χρησιμοπιώντας την X^2 κατανομή ως εξής:

$$X^2 = D(M_o) - D(M_k) = -2 \times LL(M_o) - (-2 \times LL(M_k)) = 2 \times LL(M_k) - 2 \times LL(M_o).$$

Η ποσότητα X^2 είναι πάντα θετική, αφού $D(M_o) > D(M_k)$. Μεγάλες τιμές της X^2 δηλώνουν πως το μοντέλο M_k προσαρμόζεται καλύτερο από το M_o.

3.1.5 Διαστήματα Εμπιστοσύνης και Έλεγχοι Υποθέσεων

Ως διάστημα εμπιστοσύνης (δ.ε.) ορίζουμε ένα διάστημα που περιέχει την τιμή της παραμέτρου που εξετάζουμε, με μια πιθανότητα. Η συγκεκριμένη πιθανότητα συνήθως συμβολίζεται με γ και εξαρτάται από το επίπεδο σημαντικότητας που θέτουμε κάθε φορά, a. Πιο συγκεκριμένα, $\gamma = 1 - a$, την οποία ονομάζουμε συντελεστή/επίπεδο εμπιστοσύνης. Η μορφή ενός διαστήματος εμπιστοσύνης είναι:

$$[Εκτιμητής] \pm Z_{a/2}[Τυπικό\ σφάλμα\ του\ εκτιμητή].$$

Η τιμή $z_{a/2}$ εξαρτάται από το επίπεδο σημαντικότητας a και εξάγεται από τους πίνακες της κανονικής κατανομής.

Για την δημιουργία διαστημάτων εμπιστοσύνης που αφορούν τους εκτιμητές των συντελεστών, b_j, μπορούμε να χρησιμοποιήσουμε τη στατιστική συνάρτηση του Wald. Το στατιστικό του Wald ορίζεται ως ο λόγος

του εκτιμητή προς το τυπικό του σφάλμα, $(\frac{\hat{b}_j}{se(\hat{b}_j)})^2$, το οποίο χρησιμοποιείται για τον έλεγχο υποθέσεων με μηδενική υπόθεση, $H_o : b_j = 0$ (ο συντελεστής να είναι ίσος με 0, δεν έχει επίδραση στο μοντέλο) έναντι της εναλλακτικής, $H_1 : b_j \neq 0$ (σημαντική επίδραση στο μοντέλο). Το στατιστικό ακολουθεί προσεγγιστικά την X^2 κατανομή, με ένα βαθμό ελευθερίας. Αυτό συμβαίνει γιατί το b_j, σύμφωνα με το Κεντρικό Οριακό Θεώρημα (Κ.Ο.Θ.), προσεγγίζει την Κανονική Κατανομή, λόγω του μεγάλου μεγέθους δείγματος. Άρα, ο $b_j \sim N(0, se(\hat{b}_j))$ με μέση τιμή 0 κάτω από την H_o. Αφού το στατιστικό του Wald είναι κανονικοποιημένο, θα ακολουθεί προσεγγιστικά την $N(0, 1)$. Γνωρίζουμε πως αν υψώσουμε στο τετράγωνο μια τυποποιημένη μεταβλητή που ακολουθεί την κανονική κατανομή, τότε η μεταβλητή που προκύπτει θα ακολουθεί την X^2 κατανομή, με ένα βαθμό ελευθερίας.

Επιπλέον, δημιουργούνται τα εξής διαστήματα εμπιστοσύνης για τους εκτιμητές των συντελεστών:

$$\hat{b}_j \pm Z_{a/2} se(\hat{b}_j).$$

Κάνοντας χρήση της εκθετικής συνάρτησης μπορούμε να υπολογίσουμε τα διαστήματα για τον λόγο των συμπληρωματικών πιθανοτήτων (odds),

$$exp(\hat{b}_j \pm Z_{a/2} se(\hat{b}_j)).$$

Με την χρήση του στατιστικού Wald μπορούμε να δημιουργήσουμε το $100\%(1 - a)$ διάστημα επιστοσύνης για την μέση τιμή της συνάρτησης $\hat{g(x)} = a + b_j x_j$ με $j = 1, \ldots, p$. Η μορφή του διαστήματος θα είναι:

$$\hat{g(x)} \pm Z_{a/2} se(\hat{g(x)}).$$

Το τυπικό σφάλμα της $\hat{g(x)}$ είναι ίσο με την τετραγωνική ρίζα της διασποράς της, $\sqrt{Var(\hat{g(x)})}$.

3.1.6 Υπόλοιπα και Έκτροπες Τιμές

Ο υπολογισμός των υπολοίπων χρησιμοποιείται για τον έλεγχο της συμφωνίας μεταξύ των παρατηρήσεων και των προσαρμοσμένων τιμών, την έρευση έκτροπων τιμών που επηρεάζουν αρνητικά το μοντέλο. Υπάρχουν δύο είδη υπολοίπων, τα τυποποιημένα υπόλοιπα (Deviance Residuals) και τα υπόλοιπα κατά Pearson.

Τα υπόλοιπα μέτρου απόκλισης ορίζονται ως:

$$d_i = sign(y_i - \hat{p}_i)\sqrt{-2(y_i log(\hat{p}_i) + (1 - y_i) log(1 - \hat{p}_i))},$$

και τα τυποποιημένα μέτρα απόκλισης ορίζονται ως:

$$d_i^s = \frac{d_i}{\sqrt{(1 - h_{ii})}}.$$

Τα υπόλοιπα κατά Pearson ορίζονται ως:

$$r_i^P = \frac{y_i - \hat{p}_i}{\sqrt{\hat{p}_i(1 - \hat{p}_i)}},$$

και τα τυποποιημένα υπόλοιπα Pearson ορίζονται ως:

$$r_i^{PD} = \frac{r_i^P}{\sqrt{(1 - h_{ii})}}.$$

Οι τιμές h_{ii} που εμφανίζονται στον παρονομαστή των τυποποιημένων υπολοίπων ονομάζονται μόχλευση και αποτελούν τα διαγώνια στοιχεία του πίνακα προβολής H, ο οποίος είναι διάστασης $n \times n$ και ορίζεται ως,

$$H = W^{1/2} X (X^T W X)^{-1} X^T W^{1/2}.$$

Ο X είναι ο πίνακας σχεδιασμού διάστασης $n \times p$ και ο W είναι ο διαγώνιος πίνακας διάστασης $n \times n$ με στοίχεια του,

$$\hat{p}_i(1 - \hat{p}_i),$$

που αποτελούν την εκτιμώμενη διασπορά $V(\hat{y}_i)$ της παρατήρησης y_i.

Ο πίνακας X είναι της μορφής:

$$X = \begin{pmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1p} \\ 1 & x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & x_{n2} & \cdots & x_{np} \end{pmatrix},$$

και παρατηρούμε πως περιλαμβάνει τις τιμές των παρατηρήσεων, x_{ij}, όπου ο δείκτης i αναφέρεται στον αριθμό της παρατήρησης και ο δείκτης j στην επεξηγηματική μεταβλητή που αναφέρεται. Η πρώτη στήλη που περιέχει την μονάδα αναφέρεται στον εκτιμητή της σταθεράς του μοντέλου.

Ο διαγώνιος πίνακας W έχει την μορφή:

$$W = \begin{pmatrix} p_1(1-p_1) & 0 & \cdots & 0 \\ 0 & p_2(1-p_2) & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & p_n(1-p_n) \end{pmatrix}.$$

Το άθροισμα των h_{ii} ισούται με,

$$\sum_{i=1}^{n} h_{ii} = k + 1.$$

Επομένως, μέση τιμή των h_{ii} είναι $\frac{k+1}{n}$. Οι τιμές h_{ii} αποτελούν ένα μέτρο για τον έλεγχο της επίδρασης των παρατηρούμενων πάνω στις προβλεπόμενες τιμές της μεταβλητής απόκρισης, δηλαδή την έρευση έκτροπων τιμών. Συνήθως ένα εμπειρικό κριτηρίο είναι η ανισότητα, $h_{ii} > \frac{3(k+1)}{n}$. Άρα, σημεία με υψηλή μόχλευση ασκούν υψηλή επιρροή και είναι απομακρυσμένα από τις υπόλοιπες παρατηρήσεις (*outlier*). Τα στοιχεία μόχλευσης δείχνουν πόσο μεγάλη επίδραση έχει η i παρατήρηση στο μοντέλο.

Απόσταση Cook

Η απόσταση Cook εξετάζει την επίδραση που θα έχει η αφαίρεση ενός σημείου επιρροής στην εκτίμηση του διανύσματος των παραμέτρων του μοντέλου. Εστώ, $\hat{b}$ το διάνυσμα των παραμέτρων λαμβάνοντας υπόψιν όλες τις παρατηρήσεις και $\hat{b_i}$ το διάνυσμα των παραμέτρων αν η i-οστή παρατήρηση αφαιρεθεί. Αντιλαμβανόμαστε πως θα γίνει χρήση της διαφοράς $\hat{b} - \hat{b_i}$, η οποία θα είναι αξιοσημείωτη αν η παρατήρηση i είναι σημείο επιρροής. Η απόσταση ορίζεται ως:

$$D_i = \frac{(\hat{b_i} - \hat{b})' X' W X (\hat{b_i} - \hat{b})}{k+1} = \frac{r_i^{PD^2} h_{ii}}{(1-hii)(k+1)},$$

για τιμές $D_i > 1$ θεωρούμε πως υπάρχει πρόβλημα.

Στην περίπτωση που θέλουμε να παρατηρήσουμε την επίδραση μιας παρατήρησης σε κάθε παράμετρο ξεχωριστά και όχι στο διάνυσμα των συντελεστών, μπορούμε να χρησιμοποιήσουμε το μέτρο DFBETAS. Στο μέτρο αυτό συγκρίνουμε τις εκτιμήτριες μεγίστης πιθανοφάνειας $\hat{b_j}$ για την j παράμετρο με την αντίστοιχη $\hat{b\_j}$ της παραμέτρου j αν έχουμε αφαιρέσει την i-οστή παρατήρηση. Για τιμές, $|\hat{b_j} - \hat{b_{-j}}| > 1$ καταλαβαίνουμε ότι η i-οστή παρατήρηση έχει επιρροή στην εκτίμηση της j-οστής παραμέτρου.

3.1.7 Ταξινόμηση

Ως ταξινόμηση, ορίζεται η διαδικασία μέσω της οποίας προβλέπεται η τιμή της μεταβλητής απόκρισης ενός συνόλου δεδομένων χρησιμοποιώντας τις υπάρχουσες τιμές των ανεξάρτητων μεταβλητών. Για το λόγο αυτό δημιουργείται ένας δισδιάστατος πίνακας, ο οποίος ονομάζεται πίνακας σύγχυσης. Για την δημιουργία του παραπάνω πίνακα είναι απαραίτητο να θέσουμε ένα όριο για την εκτίμηση της πιθανότητας επιτυχίας του μοντέλου, έστω $\hat{p}$. Το όριο αυτό ονομάζεται σημείο διαχωρισμού και το συμβολίζουμε με $p*$. Αν $\hat{p} > p*$, τότε το $Y = 1$, διαφορετικά το $Y = 0$. Ο πίνακας σύγχυσης έχει την παρακάτω μορφή:

Πραγματικότητα / Πρόβλεψη	$Y = 1$	$Y = O$
$Y = 1$	TP	FP
$Y = O$	FN	TN

Οι τιμές του πίνακα προκύπτουν μεταξύ της συμφωνίας της πραγματικής τιμής και της τιμής πρόβλεψης και ερμηνεύονται ως εξής:

- TP = True Positive, ορθές θετικές.
- TN = True Negative, ορθές αρνητικές.
- FP = False Positive, ψευδείς θετικές.
- FN = False Negative, ψευδείς αρνητικές.

Κάνοντας χρήση του παραπάνω πίνακα μπορούμε να υπολογίσουμε τρείς βασικές τιμές:

1. **Ευαισθησία**: η πιθανότητα η πρόβλεψη να είναι θετική, $Y = 1$.

   Ευαισθησία (sensitivity)=$\frac{TP}{TP+FN}$.

2. **Ειδικότητα**: η πιθανότητα πρόβλεψης να είναι η κατάσταση $Y = 0$.

   Ειδικότητα (specificity)=$\frac{TN}{TN+FP}$.

3. **Ακρίβεια**: η συνολικά ορθή πρόβλεψη.

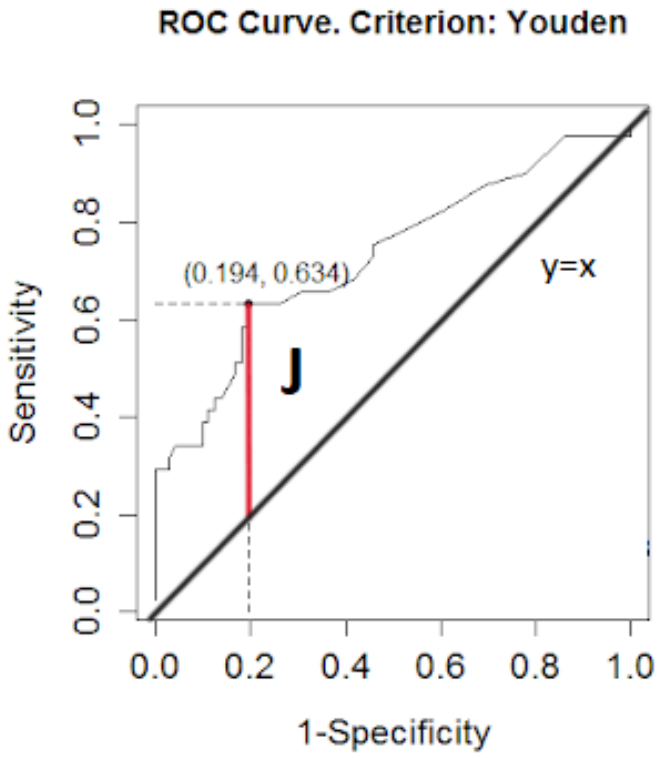
   Ακρίβεια (accuracy)=$\frac{TP+TN}{TP+TN+FP+FN}$.

Γίνεται αντιληπτό ότι η τιμή του σημείου διαχωρισμού, $p*$, επηρεάζει την μορφή του πίνακα και για αυτό το λόγο είναι σημαντικό να υπόλογίζεται η βέλτιστη τιμή του ορίου. Για την εύρεση του ορίου χρησιμοποιούμε την συνάρτηση χρησιμότητας (utility function), η οποία προκύπτει από

τον σταθμικό μέσο της ευαισθησίας και της ειδικότητας, δίνοντας βάρος ανάλογα με τον δείκτη που δίνουμε μεγαλύτερη αξία (ανάλογα το ερευνητικό ερώτημα). Επομένως, για πεδίο τιμών του $p* \in [0.01, 0.99]$ παρατηρούμε για ποια τιμή του μεγιστοποιείται η συνάρτηση χρησιμότητας.

Διαφορετικά, επιλέγουμε ως σημείο διαχωρισμού την μέγιστη τιμή του δείκτη Youden, *J=ευαισθησία+ειδικότητα-1*.

Υπολογίζοντας την ποσότητα *1-ειδικότητα* μπορούμε να δημιουργήσουμε τις καμπύλες ROC (Receiver Operating Characteristic Curve). Η καμπύλη ROC απεικονίζει την προβλεπτική ικανότητα του μοντέλου καθώς το σημείο διαχωρισμού μεταβάλλεται. Στον κατακόρυφο άξονα απεικονίζονται οι τιμές της ευαισθησίας (sensitivity) και στον οριζόντιο άξονα οι τιμές των $Y = 0$ που προβλέφθηκαν λανθασμένα, δηλαδή η συμπληρωματική πιθανότητα της ειδικότητας (*1-ειδικότητα*).

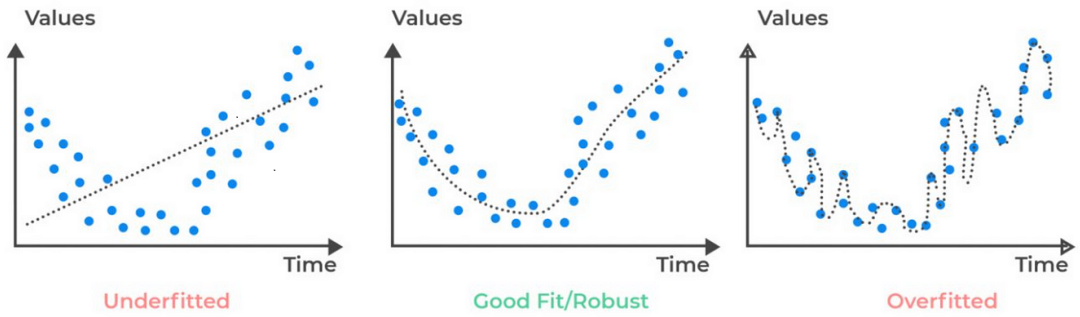


Σχήμα 3.1: Καμπύλη *ROC*, η ευθεία $y = x$ και ο δείκτης J

3.1.8 Μέθοδος Διασταυρωμένης Επικύρωσης

Για την δημιουργία του μοντέλου της λογιστικής παλινδρόμησης έχουν χρησιμοποιηθεί αποκλειστικά δεδομένα που έχουμε στην διάθεση μας,

επομένως δεν γνωρίζουμε αν το μοντέλο αυτό θα έχει την επιθυμητή ακρίβεια αν χρησιμοποιηθεί για νέο σύνολο δεδομένων. Άρα, μπορεί να οδηγηθούμε σε προβλήματα υπερπροσαρμογής (overfitting) ή υποπροσαρμογής (underfitting), όπως φαίνεται στο Σχήμα 3.2. Ως υπερπροσαρμογή ορίζεται η απόδοση του μοντέλου στα δεδομένα εκπαίδευσης, αλλά μέτρια απόδοση σε νέα δεδομένα, και ως υποπροσαρμογή, είναι η φτωχή προσαρμογή ακόμα και στα δεδομένα εκπαίδευσης. Ένα ακόμη πρόβλημα που δημιουργείται από τα παραπάνω είναι ο συμβιβασμός μεταξύ μεροληψίας και διασποράς. Υψηλή μεροληψία υπάρχει όταν το μοντέλο κάνει λάθη για κάθε σύνολο εκπαίδευσης, ενώ έχουμε ύπαρξη υψηλής διασποράς όταν δύο σύνολα εκπαίδευσης οδηγούν σε πολύ διαφορετικά μοντέλα. Η υπερπροσαρμογή συνδέεται με χαμηλή μεροληψία και υψηλή διασπορά, ενώ η υποπροσαρμογή με υψηλή μεροληψία και χαμηλή διασπορά.



Σχήμα 3.2: Μοντέλο με υποπροσαρμογή, καλή προσαρμογή και υπερπροσαρμογή

Για την επίλυση των παραπάνω προβλημάτων σε ένα μοντέλο μπορούμε να χρησιμοποιήσουμε την μέθοδο της διασταυρωμένης επικύρωσης. Με αυτό τον τρόπο μπορούμε να ελέγξουμε την αποτελεσματικότητα του μοντέλου, αποτελεί μια μέθοδο επαναδειγματοληψίας. Η ιδέα πίσω από την μέθοδο είναι ο διαχωρισμός των δεδομένων σε δύο ανεξάρτητα σύνολα, ένα σύνολο για την ανάπτυξη του μοντέλου, το οποίο ονομάζεται σύνολο εκπαίδευσης (training) και σε ένα σύνολο για την προβλεπτική ικανότητα του μοντέλου. Με αυτό τον τρόπο επιλύεται το θέμα της εντός δείγματος πρόβλεψης και οδηγούμαστε στην εκτός δείγματος.

Υπάρχουν δύο πιθανοί τρόποι να διαχωρίσουμε το σύνολο δεδομένων. Αρχικά, μπορούμε να χωρίσουμε το σύνολο σε $n - k$ παρατηρήσεις ως σύνολο εκπαίδευσης και οι υπόλοιπες παρατηρήσεις, k, για την εξέταση. Το μοντέλο προσαρμόζεται στο σύνολο εκπαίδευσης, υπολογίζουμε τους συντελεστές του και δείκτες όπως τα συνολικά σφάλματα προβλέψεων

στα δεδομένα εξέτασης. Η συγκεκριμένη διαδικασία επαναλαμβάνεται για αρκετές φορές, όπου κάθε φορά το σύνολο εκπαίδευσης και εξέτασης αλλάζει (η αναλογία παραμένει ίδια) και στο τέλος υπολογίζουμε τον μέσο όρο των δεικτών που έχουν προκύψει σε κάθε διαδικασία. Όπως γίνεται αντιληπτό, στην συγκεκριμένη μέθοδο μπορεί να υπάρξουν δεδομένα που δεν επιλέγησαν ποτέ για εκπαίδευση και αντίστιχα για εξέταση.

Μια ακόμη τεχνική διαχωρισμού των δεδομένων σε σύνολο εκπαίδευσης και εξέτασης είναι η $k-fold$. Στην συγκεκτιμένη τεχνική το σύνολο των δεδομένων χωρίζεται σε k υποσύνολο, ισοπληθικά μεταξύ τους. Το ένα εκ αυτών χρησιμοποιείται ως εκπαίδευσης, ενώ η ένωση των υπόλοιπων $k-1$ ως εξέτασης, και υπόλογίζουμε τους επιθυμητούς δείκτες. Η παραπάνω διαδικασία επαναλαμβάνεται τόσες φορές, όσες το κάθε ένα υποσύνολο να έχει χρησιμοποιηθεί ως σύνολο εκπαίδευσης και τελικά υπολογίζεται ο μέσος όρος των δεικτών. Με αυτό τον τρόπο το μειονέκτημα της προηγούμενης τεχνικής διαχωρισμού επιλύεται, αφού κάθε παρατήρηση θα χρησιμοποιηθεί και για εκπαίδευση και για εξέταση.

Τόσο η ταξνόμηση, όσο και η μέθοδος της διασταυρωμένης επικύρωσης αποτελούν εργαλεία της μηχανικής μάθησης και συγκεκριμένα της επιβλεπόμενης. Ως επιβλεπόμενη μάθηση, ορίζεται ο αλγόριθμος στον οποίο το μοντέλο εκπαιδεύεται σε δεδομένα όπου οι σωστές απαντήσεις είναι ήδη επισημασμένες (το σύνολο περιέχει και τις τιμές εισόδου και τις τιμές εξόδου). Στόχος αυτής της τεχνικής είναι το μοντέλο να εκπαιδευτεί ώστε να μπορεί να προβλέπει σωστά τις εξόδους σε νέα και άγνωστα δεδομένα. Όπως θα δούμε παρακάτω, σε μια λογιστική παλινδρόμηση μπορεί να εκπαιδευτεί για να προβλέπει την απασχόληση ενός ατόμου αν γνωρίζει χαρακτηριστικά όπως είναι η ηλικία, το φύλο και η εκπαίδευση. Αντιλαμβανόμαστε ότι η ταξινόμηση χρησιμοποιείται σε προβλήματα όπου οι μεταβλητή εξόδου είναι διαδική (εργαζόμενος/άνεργος), για να προβλέψει σε ποιά από τις δύο κατηγορίες θα ανήκει. Ενώ, η διασταυρωμένη επικύρωση χρησιμοποιείται για την αξιολόγηση της απόδοσης του μοντέλου σε μελλοντικά δεδομένα.

Κεφάλαιο 4

Λογιστική Παλινδρόμηση στα Δεδομένα της Ανεργίας

Στον παρόν κεφάλαιο, θα προσαρμοστεί και θα ερμηνευτεί ένα μοντέλο λογιστικής παλινδρόμησης με την χρήση της γλώσσας R. Στο παρόν μοντέλο, θα εξεταστεί η επίδραση διαφόρων δημογραφικών παραγόντων (ηλικία, φύλο, υπηκοότητα), καθώς και του εκπαιδευτικού επιπέδου στην πιθανότητα ανεργίας ή εργασίας. Τελικά, θα ελεχθεί η ικανότητα του μοντέλου να προσαρμοστεί σε μελλοντικά δεδομένα.

4.1 Κωδικοποίηση και Παρουσίαση Μεταβλητών

Οι μεταβλητές που χρησιμοποιούνται για την προσαρμογή του μοντέλου είναι η κατάσταση απασχόλησης, η ηλικία, το φύλο, η εκπαίδευση και η υπηκοότητα. Η μεταβλητή απόκρισης του μοντέλου θα είναι η κατάσταση απασχόλησης, η οποία αποτελείται από 3 κατηγορίες, όμως τα άτομα εκτός εργατικού δυναμικού δεν μας ενδιαφέρουν, για το λόγο αυτό εστιάζουμε στο δείγμα που αποτελείται από τις μονάδες με τιμές 1 ή 2 (δηλαδή εργασία ή ανεργία αντίστοιχα) και αφαιρούνται οι μονάδες με τιμή 3 (εκτός εργατικού δυναμικού) στην μεταβλητή που περιέχει την κατάσταση απασχόλησης. Άρα, η μεταβλητή μετατρέπεται σε δίτιμη και στην συνέχεια αλλάζουμε την κωδικοποίηση της σε 0 = άνεργος και 1 = εργαζόμενος. Οι ανεξάρητες μεταβλητές του μοντέλου θα είναι η ηλικία, η οποία είναι ποσότική, το φύλο, η εκπαίδευση και η υπηκοότητα. Οι επεξηγηματικές μεταβλητές κωδικοποιούνται ως:

1. Φύλο (gender): Άνδρας = 0, Γυναίκα = 1.
2. Εκπαίδευση (education): 0 = Χωρίς στοιχειώδη εκπαίδευση, 1 = Στοιχειώδη εκπαίδευση, 2 = Γυμνάσιο, 3 = Λύκειο, 4 = ΙΕΚ, 5 =

Προπτυχιακό, 6 = Μεταπτυχιακό.

3. Υπηκοότητα (citizenship): 0 = Ελλάδα, 1 = Ευρωπαϊκή Ένωση, 2 = Άλλη Χώρα.

Στον κώδικα που ακολουθεί, με την χρήση της βιβλιοθήκης *data.table* αφαιρούμε από το δείγμα τις παρατηρήσεις που αφορούν τα άτομα εκτός εργατικού δυναμικού, και προκύπτει το σύνολο δεδομένων dt2 το οποίο θα χρησιμοποιήσουμε. Στην συνέχεια, εξάγουμε τις μεταβλητές που θα χρησιμοποιήσουμε για το μοντέλο, ορίζουμε με την εντολή *as.factor()* τις αντίστοιχες κατηγορικές. Τέλος, κωδικοποιούμε μόνο τις κατηγορικές μεταβλητές κατάλληλα.

```
library(data.table)
dt<-setDT(df)
dt2<-dt[KATAP!=3]

#Metavlites
employment<-dt2$KATAP
employment<-as.factor(employment)
gender<-dt2$HH7
gender<-as.factor(gender)
age<-dt2$AGE
education<-dt2$Q_G1
education<-as.factor(education)
citizenship<-dt2$CITIZENSHIP
citizenship<-as.factor(citizenship)

#Kodikopoihsh
levels(employment)<-c("1","0")
levels(gender)<-c("0","1")
levels(education)<-c("0","1","2","3","4","5","6")
levels(citizenship)<-c("0","1","2","3")
```

4.2 Προσαρμογή Μοντέλου

Προσαρμόζουμε το μοντέλο της λογιστικής παλινδρόμησης στα δεδομένα μας, με την συνάρτηση *glm* και με μεταβλητή απόκρισης την employment και επεξηγηματικές μεταβλητές τις gender, age, education και citizenship. Το σύνολο δεδομένων που χρησιμοποιείται είναι το dt2 και το όρισμα family="binomial" δηλώνει πως πρόκειται για λογιστική παλινδρόμηση.

```
#prosarmogi montelou
> logit_model<-glm(employment~gender+age+education+citizenship,data=dt2,
    family = "binomial")
```

4.3 Έλεγχος Προϋποθέσων

4.3.1 Πολυσυγραμμικότητα

Αρχικά, ελέγχουμε την πολυσυγραμμικότητα του μοντέλου με την εντολή *vif()* της βιβλιοθήκης car. Ο δείκτης VIF (Variance Inflation Factor), δηλαδή ο παράγοντας μεγένθυσης διασποράς, αντιπροσωπεύει την αύξηση της διασποράς ενός συντελεστή που εκτιμάται όταν υπάρχουν συσχετίσεις μεταξύ των επεξηγηματικών μεταβλητών. Τα επιθυμητά αποτελέσματα λαμβάνονται για τιμές του GVIF μικρότερες του 5 και για τιμές του προσαρμοσμένου δείκτη (ως προς του βαθμούς ελευθερίας, df) να είναι μικρότερες του 2.

```
#Polysygrammikotita
>library(car)
> vif(logit_model)
                GVIF Df GVIF^(1/(2*Df))
gender      1.032131  1        1.015939
age         1.133565  1        1.064690
education   1.210821  6        1.016069
citizenship 1.050787  2        1.012462
```

Παρατηρώντας τα παραπάνω αποτελέσματα η τιμή του GVIF είναι μικρότερη του 5 για κάθε μεταβλητή και η τιμή του προσαρμοσμένου μικρότερη του 2 για κάθε μεταβλητή αντίστοιχα. Επομένως, συμπεραίνουμε πως δεν υπάρχει πρόβλημα πολυσυγγραμμικότητας.

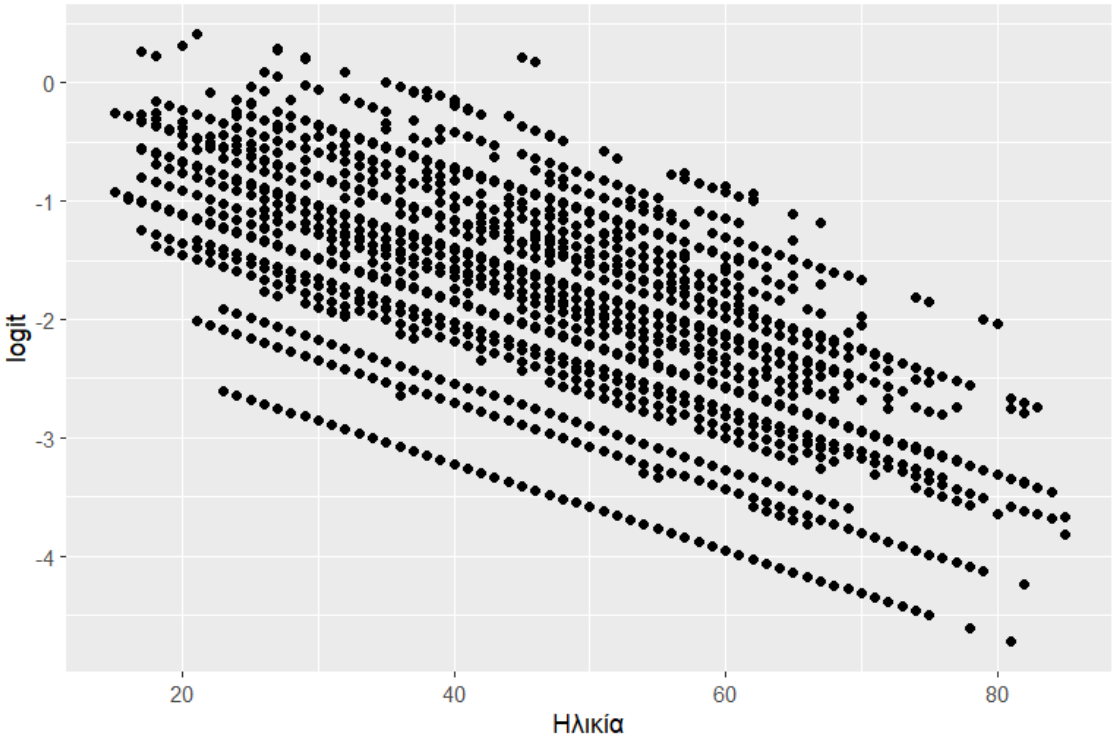
4.3.2 Ανεξαρτησία

Η προϋπόθεση της ανεξαρτησίας δεν ελέγχεται, καθώς θεωρούμε πως ισχύει από τον σχεδιασμό της έρευνας.

4.3.3 Γραμμικότητα

Ο έλεγχος της γραμμικότητας πραγματοποιείται μεταξύ των ποσοστικών επεξηγηματικών μεταβλητών και του λογαρίθμου των συμπληρωματικών πιθανοτήτων. Στην συγκεκριμένη περίπτωση, μεταξύ της μεταβλητής age και του $log(\frac{p}{1-p})$, δηλαδή της συνάρτησης *logit*. Αρχικά, υπολογίζουμε τις προσαρμοσμένες πιθανότητες, *probabilities*, τις οποίες χρησιμοποιούμε για τον υπολογισμό της *logit*. Η γραφική απεικόνιση της γραμμικότητας θα γίνει με την χρήση του πακέτου ggplot2. Δημιουργούμε ένα πλαίσιο δεδομένων με τις ήδη υπάρχουσες μεταβλητές του μοντέλου και την προσθήκη της *logit*, το οποίο θα χρησιμοποιηθεί για την γραφική απεικόνιση.

```
#Grammikotita
> library(ggplot2)
> library(cowplot)
> probabilities<-predict(logit_model,type="response")
> logit<-log(probabilities/(1-probabilities))
> length(logit)<-length(age)
> dt_test<-data.frame(logit,gender,age,education,citizenship)
> ggplot(dt_test,aes(age,logit))+
  geom_point()+ geom_smooth(method="loess")
```



Σχήμα 4.1: Έλεγχος γραμμικότητας στο λογιστικό μοντέλο παλινδρόμησης

Παρατηρώντας το Σχήμα (4.1) υπάρχει γραμμικότητα μεταξύ της ποσότικής μεταβλητής της ηλικίας και της *logit*.

4.4 Αποτελέσματα Μοντέλου και Ερμηνεία Συντελεστών

Με την εντολή *summary()* βλέπουμε τα αναλυτικά αποτελέσματα του μοντέλου.

```
> summary(logit_model)

Call:
glm(formula = employment ~ gender + age + education + citizenship,
    family = "binomial", data = dt2)

Coefficients:
```

```
               Estimate Std. Error z value Pr(>|z|)
(Intercept)    0.199545   0.149408   1.336    0.182
gender1        0.684260   0.024150  28.333  < 2e-16 ***
age           -0.036489   0.001043 -34.970  < 2e-16 ***
education1    -0.594587   0.144456  -4.116 3.85e-05 ***
education2    -0.580593   0.144619  -4.015 5.95e-05 ***
education3    -0.821090   0.141904  -5.786 7.20e-09 ***
education4    -0.919732   0.145133  -6.337 2.34e-10 ***
education5    -1.448072   0.143467 -10.093  < 2e-16 ***
education6    -1.962760   0.156367 -12.552  < 2e-16 ***
citizenship1   0.975065   0.109911   8.871  < 2e-16 ***
citizenship2   0.437677   0.054062   8.096 5.69e-16 ***
---


(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 52366  on 72518  degrees of freedom
Residual deviance: 49540  on 72508  degrees of freedom
AIC: 49562

Number of Fisher Scoring iterations: 5
```

Στον παραπάνω κώδικα παρουσιάζονται οι συντελεστές του μοντέλου. Παρατηρούμε ότι η R έχει δημιουργήσει αυτόματα εικονικές μεταβλητές για τις κατηγορικές μεταβλητές. Συγκεκριμένα, οι εικονικές μεταβλητές που δημιουργούνται είναι $m-1$, όπου m ο αριθμός των κατηγοριών της κάθε μεταβλητής. Άρα, προκύπτουν οι εξής εικονικές μεταβλητές:

- gender1, της μεταβλητής gender
- education1, education2, education3, education4, eduction5, education6, της μεταβλητής education
- citizenship1, citizenship2, της μεταβλητής citizenship

Υπολογίζουμε και τις εκθετικές τιμές των συντελεστών, οι οποίες αποτελούν τις μεταβολές του λόγου των συμπληρωματικών πιθανοτήτων (*odds*), της αντίστοιχης κατηγορίας, για παραμονή ή οχι στην κατάσταση της ανεργίας. Μπορούμε να στρογγυλοποιήσουμε τα αποτελέσματα, σε τρία δεκαδικά ψηφία με την εντολή *round*(., 3)

```
> exp(coef(logit_model))
 (Intercept)      gender1          age   education1   education2
   education3
   1.2208473    1.9823045    0.9641691    0.5517906    0.5595666
   0.4399521
   education4   education5
   0.3986257    0.2350231
  education6 citizenship1 citizenship2
   0.1404702    2.6513384    1.5491044

> round(exp(coef(logit_model)),3)
```

```
 (Intercept)      gender1          age   education1   education2
   education3
       1.221        1.982        0.964        0.552        0.560
   0.440
education4   education5
    0.399        0.235
  education6 citizenship1 citizenship2
       0.140        2.651        1.549
```

Εστιάζοντας στα αποτελέσματα, η πρώτη στήλη αποτελεί την εκτίμηση των συντελεστών, η δεύτερη το τυπικό σφάλμα, η τρίτη τον στατιστικό έλεγχο του Wald: $H_o : b_j = 0$ έναντι $H_1 : b_j \neq 0$ και η τέταρτη την P-τιμή ($P - value$), από την τιμή της οποίας καθορίζουμε το αν θα απορρίψουμε την μηδενική υπόθεση ή όχι του ελέγχου Wald. Το επίπεδο σημαντικότητας που χρησιμοποιείται είναι το 5%.

Ονοματοδοτούμε τους συντελεστές με b_j με $j = 1, ..., 10$ και τον σταθερό όρο με a. Ο συντελεστής b_1 αναφέρεται στην επίδραση του γυναικείου φύλου σε σχέση με το ανδρικό, που αποτελεί την κατηγορία αναφοράς. Ο συντελεστής b_2 αναφέρεται στην επίδραση της ηλικίας. Οι συντελετές b_3 έως b_8 αναφέρονται στην επίδραση του ανώνατου επιπέδου εκπαίδευσης σε σχέση με το επίπεδο εκπαίδευσης που αποτελεί την κατηγορία αναφοράς, το οποίο είναι "χωρίς στοιχειώδη εκπαίδευση". Τέλος, οι συντελεστές b_9 και b_{10} αναφέρονται στην επίδραση της υπηκοότητας σε σχέση με την Ελληνική υπηκοότητα, που αποτελεί την κατηγορία αναφοράς.

Ο στατιστικός έλεγχος του $Wald$, που πραγματοποιείται στην τρίτη στήλη, εξετάζει κάθε συντελεστή ξεχωριστά, αν είναι στατιστικά σημαντικός. Η μηδενική υπόθεση $H_o : b_j = 0$, σημαίνει ότι δεν υπάρχει σημαντική επίδραση της μεταβλητής στο αποτέλεσμα, ενώ η εναλλακτική υπόθεση $H_1 : b_j \neq 0$, υπάρχει στατιστικά σημαντική επίδραση στο αποτέλεσμα.

Σταθερός Όρος

Ο εκτιμητής του σταθερού όρου ισούται με $a = 0.199$, με τυπικό σφάλμα 0.15, στατιστικό έλεγχο $Wald = 1.336$ και $P - value = 0.182 > 0.05$, άρα δεν έχουμε σοβαρές ενδείξεις για να απορρίψουμε την μηδενική υπόθεση, $H_o : a = 0$, η οποία θέλει τον σταθερό όρο να είναι ίσος με 0.

Ο λόγος των συμπληρωματικών πιθανοτήτων για ένα άνδρα ηλικίας μηδέν ετών χωρίς στοιχειώδη εκπαίδευση και με ελληνική υπηκοότητα να εξέλθει από την κατάσταση της ανεργίας αυξάνεται, αφού $exp(0.199) = 1.22$. Η πιθανότητα να βρεί ένα τέτοιο άτομο εργασία είναι 22% υψηλότερη από

το να παραμείνει άνεργο.

Φύλο

Ο συντελεστής $b_1 = 0.68$ έχει τυπικό σφάλμα 0.024, στατιστικό έλεγχο $Wald = 28.33$ και $P - value < 0.05$, άρα έχουμε σοβαρές ενδείξεις για να απορρίψουμε την μηδενική υπόθεση που θέλει τον συντελεστή για το γυναικείο φύλο να είναι μηδέν, $H_o : gender1 = 0$ έναντι της $H_1 : gender1 \neq 0$.

Η ποσότητα $exp(0.68) = 1.982$, δηλώνει την αύξηση του λόγου των συμπληρωματικών πιθανοτήτων *odds* για τις γυναίκες σε σχέση με τους άνδρες για να εξέλθουν από την κατάσταση της ανεργίας, όταν οι υπόλοιπες μεταβλητές παραμένουν σταθερές. Ο λόγος των συμπληρωματικών πιθανοτήτων αυξάνεται περίπου κατα 98% για μια γυναίκα να βρεί εργασία, σε σχέση με έναν άνδρα, όταν οι υπόλοιπες μεταβλητές είναι σταθερές.

Ηλικία

Ο συντελεστής $b_2 = -0.03$ έχει τυπικό σφάλμα 0.001, στατιστικό έλεγχο $Wald = -34.97$ και $P - value < 0.05$, άρα έχουμε σοβαρές ενδείξεις για να απορρίψουμε την μηδενική υπόθεση, η οποία θέλει τον συντελεστή της ηλικίας ίσο με το 0, $H_o : age = 0$ έναντι της $H_1 : age \neq 0$.

Ο λόγος των συμπληρωματικών πιθανοτήτων μειώνεται, αφού η ποσότητα $exp(-0.036) = 0.964 < 1$, για κάθε ένα έτος που αυξάνεται η ηλικία. Τα πιο νεαρά άτομα ίδιου φύλου, επιπέδου εκπαίδευσης και υπηκοότητας έχουν μεγαλύτερο λόγο συμπληρωματικών πιθανοτήτων εύρεσης εργασίας, κατά 4%, να βρούν εργασία σε σχέση με τα μεγαλύτερα, κατά ένα έτος άτομα, με τις υπόλοιπες μεταβλητές σταθερές.

Εκπαίδευση

Ο συντελεστής $b_3 = -0.59$ έχει τυπικό σφάλμα 0.144, στατιστικό έλεγχο $Wald = -4.11$ και $P - value < 0.05$, άρα έχουμε σοβαρές ενδείξεις για να απορρίψουμε την μηδενική υπόθεση, η οποία θέλει τον συντελεστή για την στοιχειώδη εκπαίδευση να είναι ίσος με 0, $H_o : education1 = 0$ έναντι της $H_1 : education1 \neq 0$.

Ο λόγος των συμπληρωματικών πιθανοτήτων εύρεσης εργασίας μειώνεται κατα 45%, αφού $exp(-0.59) = 0.55 < 1$, για τα άτομα με στοιχειώδη εκπαίδευση σε σχέση με τα άτομα χωρίς στοιχειώδη εκπαίδευση, και τις υπόλοιπες μεταβλητές να παραμένουν σταθερές.

Ο συντελεστής $b_4 = -0.58$ έχει τυπικό σφάλμα 0.144, στατιστικό έλεγχο $Wald = -4.015$ και $P - value < 0.05$, άρα έχουμε σοβαρές ενδείξεις για να απορρίψουμε την μηδενική υπόθεση, η οποία θέλει τον συντελεστή για το γυμνάσιο να είναι ίσος με 0, H_o : $education2 = 0$ έναντι της H_1 : $education2 \neq 0$.

Ο λόγος των συμπληρωματικών πιθανοτήτων εύρεσης εργασίας μειώνεται κατα 44%, αφού $exp(-0.58) = 0.56 < 1$ για τα άτομα που έχουν ολοκληρώσει το γυμνάσιο σε σχέση με τα άτομα χωρίς στοιχειώδη εκπαίδευση, και τις υπόλοιπες μεταβλητές να παραμένουν σταθερές.

Ο συντελεστής $b_5 = -0.82$ έχει τυπικό σφάλμα 0.141, στατιστικό έλεγχο $Wald = -5.79$ και $P - value < 0.05$, άρα έχουμε σοβαρές ενδείξεις για να απορρίψουμε την μηδενική υπόθεση, η οποία θέλει τον συντελεστή για το λύκειο να είναι ίσος με 0, H_o : $education3 = 0$ έναντι της H_1 : $education3 \neq 0$.

Ο λόγος συμπληρωματικών πιθανοτήτων εύρεσης εργασίας μειώνεται κατα 56%, αφού η ποσότητα $exp(-0.82) = 0.440 < 1$ για τα άτομα που έχουν ολοκληρώσει το λύκειο σε σχέση με τα άτομα χωρίς στοιχειώδη εκπαίδευση, και τις υπόλοιπες μεταβλητές να παραμένουν σταθερές.

Ο συντελεστής $b_6 = -0.92$ έχει τυπικό σφάλμα 0.145, στατιστικό έλεγχο $Wald = -6.34$ και $P - value < 0.05$, άρα έχουμε σοβαρές ενδείξεις για να απορρίψουμε την μηδενική υπόθεση, η οποία θέλει τον συντελεστή για προπτυχιακά προγράμματα να είναι ίσος με 0,H_o : $education4 = 0$ έναντι της H_1 : $education4 \neq 0$.

Ο λόγος των συμπληρωματικών πιθανοτήτων εύρεσης εργασίας μειώνεται κατα 61%, αφού $exp(-0.92) = 0.399 < 1$, για τα άτομα που έχουν σπουδές σε ΙΕΚ σε σχέση με τα άτομα χωρίς στοιχειώδη εκπαίδευση, και τις υπόλοιπες μεταβλητές να παραμένουν σταθερές.

Ο συντελεστής $b_7 = -1.44$ έχει τυπικό σφάλμα 0.143, στατιστικό έλεγχο $Wald = -10.1$ και $P - value < 0.05$, άρα έχουμε σοβαρές ενδείξεις για να απορρίψουμε την μηδενική υπόθεση, η οποία θέλει τον συντελεστή για μεταπτυχιακά προγράμματα να είναι ίσος με 0, H_o : $education5 = 0$ έναντι της H_1 : $education5 \neq 0$.

Ο λόγος των συμπληρωματικών πιθανοτήτων εύρεσης εργασίας μειώνεται κατα 77%, αφού $exp(-0.036489) = 0.235 < 1$, για τα άτομα που έχουν προπτυχιακές σπουδές σε σχέση με τα άτομα χωρίς στοιχειώδη εκπαίδευση, και τις υπόλοιπες μεταβλητές να παραμένουν σταθερές.

Ο συντελεστής $b_8 = -1.96$έχει τυπικό σφάλμα 0.156, στατιστικό έλεγχο $Wald = -12.55$ και $P - value < 0.05$, άρα έχουμε σοβαρές ενδείξεις για να απορρίψουμε την μηδενική υπόθεση, $H_o : education6 = 0$ έναντι της $H_1 : education6 \neq 0$.

Ο λόγος των συμπληρωματικών πιθανοτήτων εύρεσης εργασίας μειώνεται κατα 86%, αφού $exp(-1.96) = 0.140 < 1$, για τα άτομα που έχουν μεταπτυχιακές σπουδές σε σχέση με τα άτομα χωρίς στοιχειώδη εκπαίδευση, και τις υπόλοιπες μεταβλητές να παραμένουν σταθερές.

Υπηκοότητα

Ο συντελεστής $b_9 = 0.975$ έχει τυπικό σφάλμα 0.10, στατιστικό έλεγχο $Wald = 8,87$ και $P - value < 0.05$, άρα έχουμε σοβαρές ενδείξεις για να απορρίψουμε την μηδενική υπόθεση, η οποία θέλει τον συντελεστή για την υπηκοότητα από χώρα της Ευρωπαϊκής Ένωσης να είναι ίσος με 0, $H_o : citizenship1 = 0$ έναντι της $H_1 : citizenship1 \neq 0$.

Ο λόγος των συμπληρωματικών πιθανοτήτων εύρεσης εργασίας αυξάνεται κατα 165%, αφού $exp(0.975) = 2.651 > 1$ για τα άτομα με Ευρωπαϊκή υπηκόοτητα σε σχέση με τα άτομα με Ελληνική υπηκοότητα, με τις υπόλοιπες μεταβλητές σταθερές. Επομένως, τα άτομα με υπηκοότητα από Ευρωπαϊκή Χώρα είναι πιο πιθανό να βρούν απασχόληση σε σχέση με τα άτομα Ελληνικής υπηκοότητας.

Ο συντελεστής $b_{10} = 0.437$έχει τυπικό σφάλμα 0.05, στατιστικό έλεγχο $Wald = 8.096$ και $P - value < 0.05$, άρα έχουμε σοβαρές ενδείξεις για να απορρίψουμε την μηδενική υπόθεση, η οποία θέλει τον συντελεστή για την υπηκοότητα από Άλλη Χώρα να είναι ίσος με 0, $H_o : citizenship2 = 0$ έναντι της $H_1 : citizenship2 \neq 0$..

Ο λόγος των συμπληρωματικών πιθανοτήτων εύρεσης εργασίας αυξάνεται κατα 54%, αφού $exp(0.437) = 1.549 > 1$, για τα άτομα με υπηκόοτητα από Άλλη Χώρα σε σχέση με τα άτομα με ελληνική υπηκοότητα, με τις υπόλοιπες μεταβλητές σταθερές.

Σύμφωνα με τα παραπάνω αποτελέσματα καταλήγουμε στα εξής:

- Οι γυναίκες έχουν υψηλότερο λόγο συμπληρωματικών πιθανοτήτων να βρούν εργασία σε σχέση με τους άνδρες, ίσης ηλικίας, επιπέδου εκπαίδευσης και υπηκόοτητας.
- Η ηλικία έχει αρνητική επίδραση στον λόγο των συμπληρωματικών πιθανοτήτων στην εύρεση εργασίας.

- Τα άτομα με υψηλότερα επίπεδα εκπαίδευσης έχουν μικρότερο λόγο συμπληρωματικών πιθανότητων σε σχέση με τα μικρότερα επίπεδα εκπαίδευσης.
- Τα άτομα υπηκοοτήτων εκτός της Ελληνικής, έχουν υψηλότερο λόγο συμπληρωματικών πιθανοτήτων για την εύρεση εργασίας σε σχέση με τα άτομα της Ελληνικής υπηκοότητας.

Επομένως, παρατηρούμε ότι όταν ένας συντελεστής b_j είναι θετικός η ποσότητα $exp(b_j)$ είναι μεγαλύτερη της μονάδας και αυξάνει τον λόγο συμπληρωματικών πιθανοτήτων, ενώ στην αντίθετη περίπτωση ο λόγος μειώνεται.

Υπολογίζουμε και τα 95% διαστήματα εμπιστοσύνης για την εκθετική τιμή του κάθε συντελεστή που εκτιμήσαμε, τα οποία παρέχουν το εύρος τιμών του λόγου των συμπληρωματικών πιθανοτήτων που περιέχουν την πραγματική τιμή:

```
> exp(confint(logit_model))
                   2.5 %    97.5 %
(Intercept)  0.9062130 1.6288561
gender1      1.8907509 2.0784973
age          0.9621980 0.9661417
education1   0.4178310 0.7366002
education2   0.4235656 0.7471872
education3   0.3348700 0.5844841
education4   0.3014187 0.5327877
education5   0.1783177 0.3131455
education6   0.1037816 0.1916999
citizenship1 2.1312401 3.2802014
citizenship2 1.3921696 1.7208565
```

Θέλουμε τα διαστήματα εμπιστοσύνης να μην περιλαμβάνουν την τιμή 1, γιατί αν ο λόγος των συμπληρωματικών πιθανοτήτων ήταν ίσος με 1 σημαίνει ότι η επίδραση των συγκρινόμενων κατηγοριών είναι ίδια (εργασία/ανεργία). Επομένως, η μεταβλητή που μελετάται δεν έχει καμία επίδραση στο μοντέλο. Παρατηρώντας τα παραπάνω αποτελέσματα, το μόνο διάστημα εμπιστοσύνης που περιλαμβάνει την τιμή 1 είναι του σταθερού όρου. Οπότε, μπορούμε να θεωρήσουμε ότι η επίδραση αυτής της μεταβλητής δεν είναι στατιστικά σημαντική.

Με την χρήση της βιβλιοθήκης *aod* θα ελέγξουμε την συνολική επίδραση του φύλου (μεταβλητές 2:2), του εκπαιδευτικού επιπέδου (μεταβλητές 4:9), καθώς και της υπηκοότητας (μεταβλητές 10:11).

```
> library(aod)
> wald.test(b=coef(logit_model),Sigma = vcov(logit_model),Terms = 2:2 )
Wald test:
----------
```

```
5
6 Chi-squared test:
7 X2 = 802.8, df = 1, P(> X2) = 0.0
8 > wald.test(b=coef(logit_model),Sigma = vcov(logit_model),Terms = 4:9)
9 Wald test:
10 ----------
11
12 Chi-squared test:
13 X2 = 828.1, df = 6, P(> X2) = 0.0
14 > wald.test(b=coef(logit_model),Sigma = vcov(logit_model),Terms = 10:11)
15 Wald test:
16 ----------
17
18 Chi-squared test:
19 X2 = 139.5, df = 2, P(> X2) = 0.0
```

Και στους τρείς ελέγχους, η $P - value$ έχει μηδενική τιμή και άρα είναι μικρότερη του επιπέδου σημαντικότητας, 0.05. Άρα, υπάρχουν σοβαρές ενδείξεις ότι η επιδράση της κάθε μεταβλητής ξεχωριστά στο μοντέλο είναι στατιστικά σημαντική.

4.5 Προβλεπόμενες Τιμές

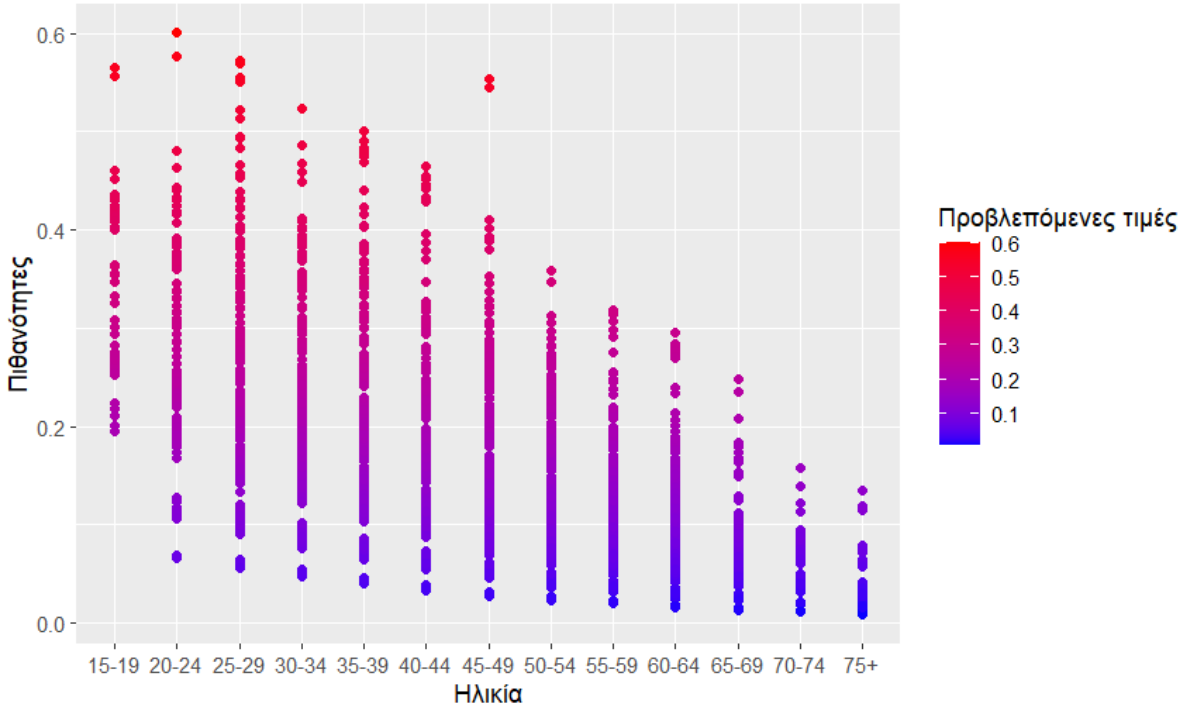
Υπολογίζουμε τις προβλεπόμενες πιθανότητες του μοντέλου με την εντολή *predict*. Στην συνέχεια, απεικονίζουμε τις τιμές αυτές σε σχέση με την ηλικία. Η ηλικία ομαδοποιείται ανά 5ετείς ομάδες.

```
1 library(ggplot2)
2
3 predicted_probabilities<- predict(logit_model, type = "response")
4
5 dataAge<- data.frame(age = dt2$age, predicted_probs =
     predicted_probabilities)
6
7 ggplot(dataAge, aes(x = age, y = predicted_probabilities, color =
     predicted_probabilities)) +
8   geom_point() +
9   scale_color_gradient(low = "blue", high = "red")
```

Στο διάγραμμα που δημιουργείται με τον παραπάνω κώδικα, το μπλέ χρώμα χρησιμοποιείται για τις χαμηλές πιθανότητες απασχόλησης, ενώ το κόκκινο για υψηλές πιθανότητες απασχόλησης.

Σύμφωνα με το Σχήμα 4.2 παρατηρούμε ότι οι ηλικιακές ομάδες που έχουν υψηλότερη πιθανότητα απασχόλησης είναι εκείνες με τα περισσότερα κόκκινα σημεία, όπως είναι οι 25-29, 30-34. Οι ηλικιακές ομάδες με τα περισσότερα μπλέ σημεία έχουν χαμηλή πιθανότητα να βρούν εργασίας, όπως είναι τα άτομα 75+. Γενικότερα, παρατηρούμε ότι όσο αυξάνεται η ηλικία οι πιθανότητα για εύρεση απασχόλησης μειώνε-

ται, κάτι το οποίο επιβεβαίωνεται και από τον συντελεστή της ηλικίας στο μοντέλο που προσαρμόσαμε.

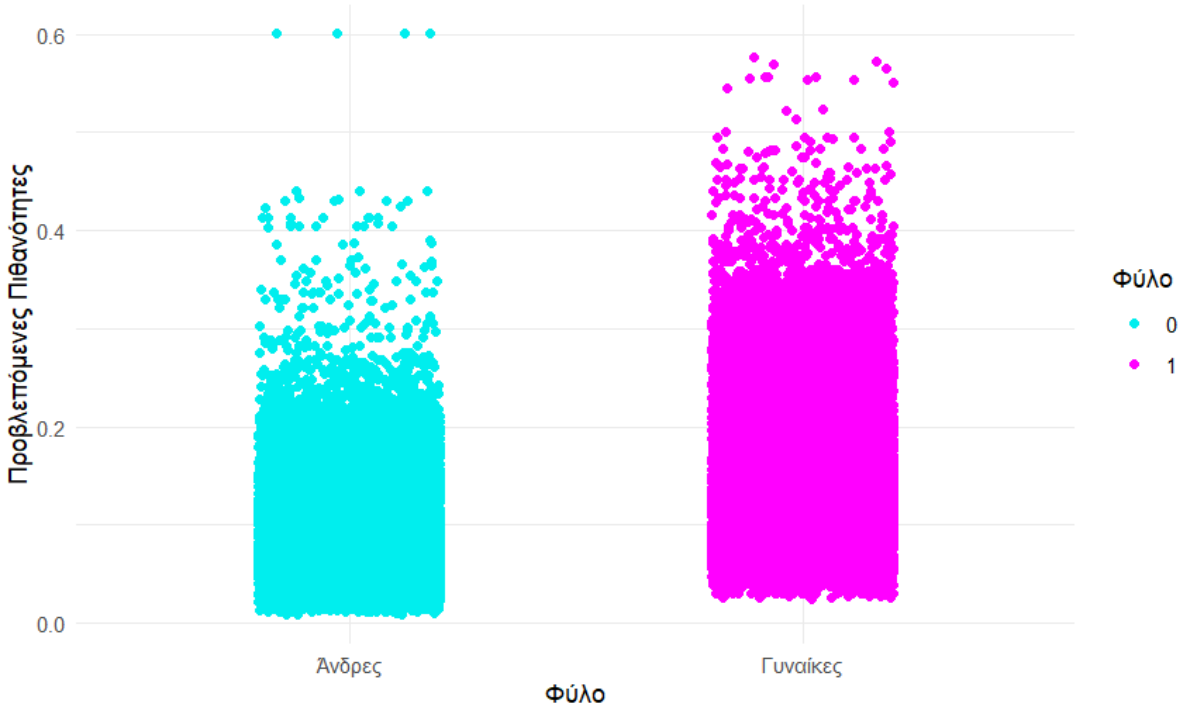


Σχήμα 4.2: Διάγραμμα Προβλεπόμενων Πιθανοτήτων ανά 5ετή ηλικιακή ομάδα

Ομοίως τις προβλεπόμενες τιμές για την απασχόληση ανάμεσα στα δύο φύλα προκύπτουν ως:

```
ggplot(dt2, aes(x = as.factor(gender), y = predicted_probabilities, color
    = as.factor(gender))) +
  geom_jitter(width = 0.2) +
  scale_color_manual(values = c("cyan2", "magenta")) +
  theme_minimal()
```

Στο Σχήμα 4.3 παρατηρούμε ότι οι προβλεπόμενες πιθανότητες των ανδρών είναι υψηλότερες, παρόλα αυτά αφορούν μόνο 4 μεμονωμένες παρατηρήσεις. Αντίθετα, οι παρατηρήσεις των γυναικών φαίνεται να έχουν πιο υψηλές προβλεπόμενες τιμές σε σχέση με τους άνδρες. Αυτό φαίνεται από το σύνολο των μεμονωμένων παρατηρήσεων, των οποίων οι προβλεπόμενες πιθανότητες είναι πιο υψηλές, δηλαδή πιο κοντά στην μονάδα. Επομένως, επιβεβαιώνεται και από εδώ ότι οι γυναίκες είναι πιο πιθανό να βρούν εργασία σε σχέση με τους άνδρες.

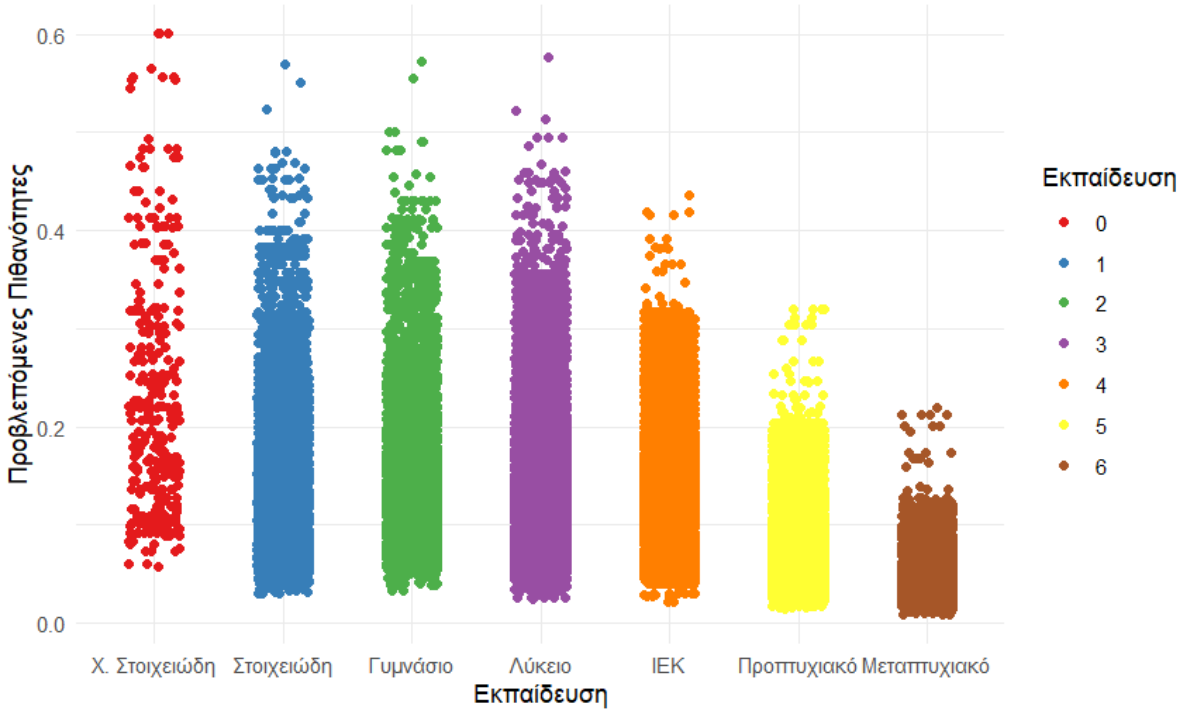


Σχήμα 4.3: Διάγραμμα προβλεπόμενων πιθανοτήτων ανά Φύλο

Οι προβλεπόμενες πιθανότητες ανά ανώτατο επίπεδο εκπαίδευσης απεικονίζονται ως εξής:

```
ggplot(dt2, aes(x = as.factor(education), y = predicted_probabilities,
    color = as.factor(education))) +
  geom_jitter(width = 0.2) +
  scale_color_brewer(palette = "Set1") +
  theme_minimal()
```

Στο Σχήμα 4.4 παρατηρούμε πως τα άτομα χωρίς στοιχειώδη εκπαίδευση έχουν πιο υψηλές προβλεπόμενες τιμές, ακολουθούν τα άτομα που έχουν ολοκληρώσει το λύκειο, το γυμνάσιο και την στοιχειώδη εκπαίδευση. Στην τελευταία θέση είναι τα άτομα με μεταπτυχιακό τίτλο. Παρατηρούμε ότι οι παρατηρήσεις που αποτελούνται από τα άτομα χωρίς στοιχειώδη εκπαίδευση είναι λιγότερες από των υπολοίπων επιπέδων εκπαίδευσης. Συμπερασματικά, όσο πιο χαμηλά βρίσκεται ένα άτομα σε επίπεδο εκπαίδευσης, τόσο πιο πιθανό είναι να βρεί εργασία.

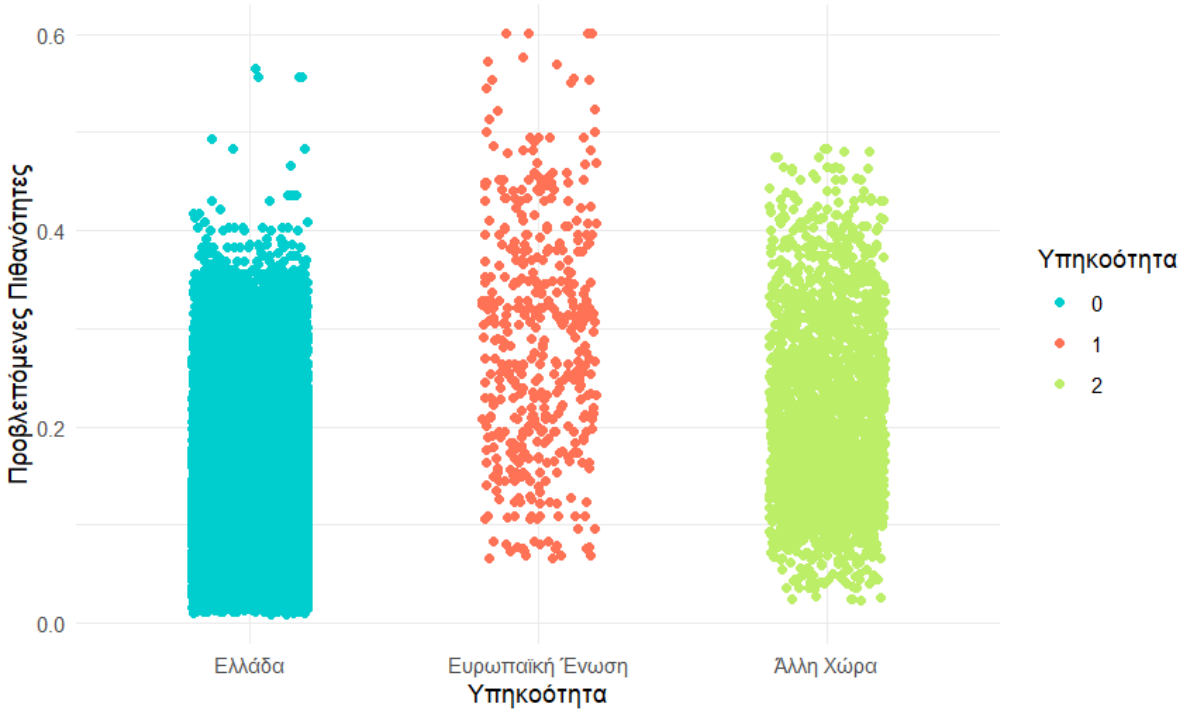


Σχήμα 4.4: Διάγραμμα προβλεπόμενων πιθανοτήτων ανά Ανώτατο Επίπεδο Εκπαίδευσης

Τέλος, υπολογίζουμε τις προβλεπόμενες τιμές για την υπηκοότητα:

```
ggplot(dt2, aes(x = as.factor(citizenship), y = predicted_probabilities,
    color = as.factor(citizenship))) +
  geom_jitter(width = 0.2) +
  scale_color_manual(values = c("cyan3", "coral1", "darkolivegreen2")) +
  theme_minimal()
```

Σύμφωνα με το Σχήμα 4.5 παρατηρούμε ότι τα άτομα με υπηκόοτητα από Ευρωπαϊκή Χώρα έχουν υψηλότερες προβλεπόμενες πιθανότητες. Κάποιες μεμονωμένες παρατηρήσεις από άτομα Ελληνικής υπηκοότητας έχουν επίσης υψηλές προβλεπόμενες πιθανότητες, όμως ο μέσος όρος αυτών των ατόμων έχει χαμηλές προβλεπόμενες πιθανότητες. Τα άτομα με υπηκόοτητα από Άλλη Χώρα έχουν υψηλότερες προβλεπόμενες πιθανότητες. Οπότε, τα άτομα με υπηκοότητα εκτός της Ελληνικής είναι πιο πιθανό να βρούν εργασία.



Σχήμα 4.5: Διάγραμμα προβλεπόμενων πιθανοτήτων ανά Υπηκοότητα

4.6 Έλεγχοι Καλής Προσαρμογής

Ελέγχουμε την καταλληλότητα του μοντέλου υπολογίζοντας το X^2, μεταξύ του μοντέλου που έχουμε αναπτύξει και του μοντέλου που περιέχει μόνο τον σταθερό όρο.

```
> mylogit_chi_sq<-logit_model$null.deviance-logit_model$deviance
> mylogit_chi_sq_df<-logit_model$df.null-logit_model$df.residual
> pchisq(mylogit_chi_sq,mylogit_chi_sq_df,lower.tail = FALSE)
[1] 0
```

Στο ίδιο αποτέλεσμα μπορούμε να καταλήξουμε χρησιμοποιώντας το πακέτο *anova*() με ορίσματα το μοντέλο που έχουμε αναπτύξει και το μοντέλο με τον σταθερό όρο.

```
> mylogit_null<-glm(employment~1,data=dt2,family = "binomial")
> anova(mylogit_null,logit_model,test="Chisq")
Analysis of Deviance Table

Model 1: employment ~ 1
Model 2: employment ~ gender + age + education + citizenship
  Resid. Df Resid. Dev Df Deviance  Pr(>Chi)
1     72518      52366
2     72508      49540 10   2826.2 < 2.2e-16 ***
---
```

Και με τους δύο τρόπους, η P-τιμή που προκύπτει είναι μικρότερη του 0.05. Έχουμε σοβαρές ενδείξεις να απορρίψουμε την μηδενική υπόθεση

που θέλει τα δυό μοντέλα, εκείνο με τον σταθερό όρο και αυτό με τις μεταβλητές της ηλικίας, του φύλου, της εκπαίδευσης και της υπηκοότητας, να έχουν την ίδια προσαρμογή.

Στην συνέχεια υπολογίζουμε τους τρείς συντελεστές προσδιορισμού, του Lemehow, Cox-Snell και Nagelkerke.

```
> R2_L<-mylogit_chi_sq/logit_model$null.deviance
> R2_L
[1] 0.0539706
>
> R2_CS<-1-exp((-mylogit_chi_sq)/nrow(dt2))
> R2_CS
[1] 0.03822272
>
> R2_N<--R2_CS/(1-exp(logit_model$null.deviance/nrow(dt2)))
> R2_N
[1] 0.03610141
```

Οι τιμές και των τριών συντελεστών είναι αρκετά χαμηλές, 5,3%, 3,8% και 3,6% αντίστοιχα. Αυτό σημαίνει ότι το μοντέλο εξηγεί ένα μικρό ποσοστό της διασποράς της μεταβλητής απόκρισης.

Η μέγιστη τιμή της συνάρτησης πιθανοφάνειας του μοντέλου είναι:

```
> logLik(logit_model)
'log Lik.' -24770 (df=11)
```

Η τιμή που προκύπτει είναι αρνητική, όπως αναμένουμε. Οι αρνητικές τιμές υποδηλώνουν καλύτερη προσαρμογή.

Τέλος, οι δείκτες AIC και BIC προκύπτουν ως:

```
> AIC(logit_model)
[1] 49562
> BIC(logit_model)
[1] 49663.1
```

Οι οποίοι προκύπτουν να έχουν παρόμοιες τιμές, με τον BIC να είναι μεγαλύτερος.

4.7 Υπόλοιπα και Έκτροπες Τιμές

Αρχικά, υπολογίζουμε τα υπόλοιπα του μέτρου απόκλισης:

```
> dev_res<-resid(logit_model)
> which(abs(as.vector(dev_res))>2)
> which(abs(as.vector(dev_res))>2)
   [1]    13    85   103   104   138   142   155   177   279   332   353
      385   400   452   585   588
 .
 .
```

```
 .
 [993] 16757 16758 16765 16766 16775 16793 16795 16803
 [ reached getOption("max.print") -- omitted 3000 entries ]
```

Από τα αποτελέσματα που προκύπτουν, όπου αναζητούμε ποια δεδομένα έχουν υπόλοιπα > 2, βλέπουμε ότι 993 παρατηρήσεις είναι εκτός των ορίων που ζητήσαμε και 3000 παρατηρήσεις που δεν εκτυπώθηκαν λόγω μεγάλου μεγέθους. Οπότε, συνολικά υπάρχουν περίπου 3993 δεδομένα των οποίων οι παρατηρούμενες με τις προβλεπόμενες τιμές έχουν διαφορά μεγαλύτερη του 2.

Εν συνεχεία, υπολογίζουμε τα υπόλοιπα Pearson:

```
> #Pearson
> pr_res<-residuals(logit_model,"pearson")
> leverages<-hatvalues(logit_model)
> stand_res<-pr_res/sqrt(1-leverages)
> summary(stand_res)
     Min.   1st Qu.    Median      Mean   3rd Qu.      Max.
-1.168311 -0.399592 -0.312615  0.000751 -0.244451  8.503707
```

Ο μέσος όρος των τυποποιημένων υπολοίπων Pearson (κανονικοποιημένα υπόλοιπα) είναι 0. Υπάρχει ένα μεγάλο εύρος τιμών που κυμαίνεται από το -1.16 έως το 8.50.

Και τέλος, για να υπολογίσουμε τα τυποποιημένα υπόλοιπα κατά Student:

```
> rst_res<-rstudent(logit_model)
> summary(rst_res)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
-1.3123 -0.5443 -0.4318 -0.1855 -0.3407  2.9317
> which(abs(as.vector(rst_res))>2)
> which(abs(as.vector(rst_res))>2)
   [1]    13    85   103   104   138   142   155   177   279   332   353
     385   400   452   585   588
 .
 .
 .
 [993] 16750 16757 16758 16765 16766 16775 16793 16795
 [ reached getOption("max.print") -- omitted 3004 entries ]
> length(a)/nrow(dt2)*100
[1] 5.521312
```

Η μέγιστη τιμή, 2.93 και η ελάχιστη τιμή -1.31 έχουν μικρότερο εύρος και η μέση τιμή παραμένει κοντά στο 0. Υπάρχουν περίπου 3.997 δεδομένα με υπόλοιπα που έχουν τιμή > 2. Το 5.52% των παρατηρήσεων έχουν τυποποιημένα υπόλοιπα Student μεγαλύτερα του 2.

Παρατηρούμε ότι για τον υπολογισμό των τυποποιημένων υπολοίπων χρησιμοποιούμε τις τιμές μόχλευσης, h_{ii}, τις οποίες υπολογίζουμε με την

εντολή *hatvalues*() και σαν όρισμα το μοντέλο μας, logit_model.

```
> leverages<-hatvalues(logit_model)
> summary(leverages)
     Min.   1st Qu.    Median      Mean   3rd Qu.      Max.
4.062e-05 6.028e-05 8.610e-05 1.517e-04 1.522e-04 7.709e-03
> s<-10/nrow(dt2)
>0.0001378949
> which(as.vector(leverages)>3*s)
> which(as.vector(leverages)>3*s)
 which(as.vector(leverages)>3*s)
   [1]     8   166   514   524   717   778   893   988  1002  1003  1029
    1058  1098  1099
  .
  .
  .
 [995] 30061 30089 30115 30144 30155 30158
 [ reached getOption("max.print") -- omitted 1984 entries ]
 > length(which(as.vector(leverages) > 3 * s))
[1] 2984

```

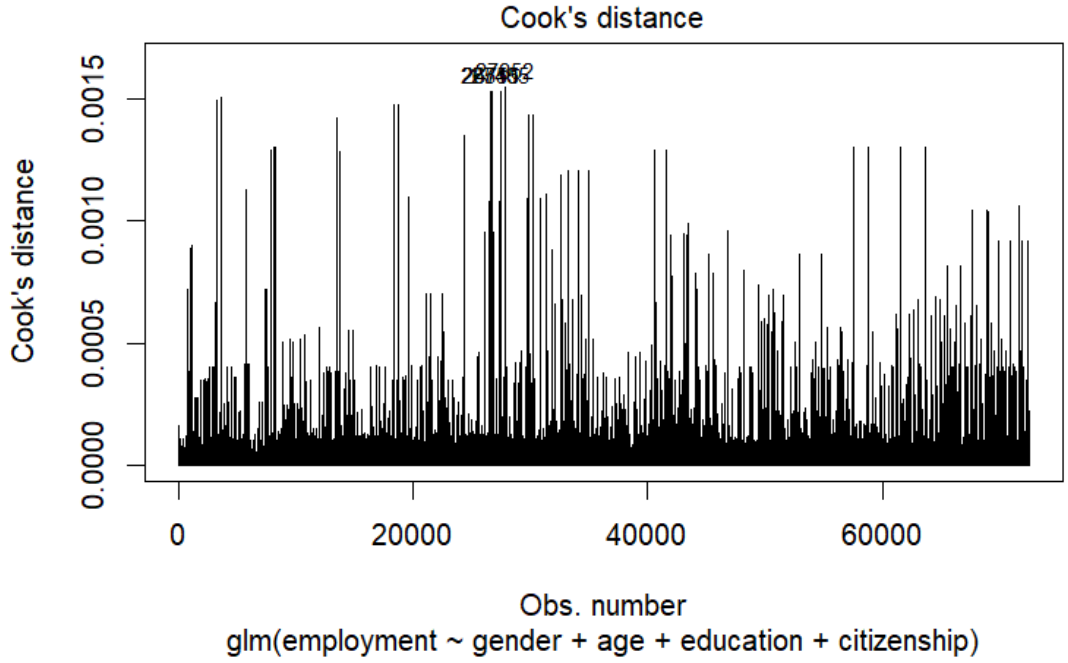
Η εντολή which μας δείχνει ότι 995 παρατηρήσεις έχουν τιμή μόχλευσης μεγαλύτερη του $3*(p+1)/n$. Επιπλέον, δηλώνει πως λόγω του μεγάλου μεγέθους του αποτελέσματος έχουν παραληφθεί 1984 παρατηρήσεις. Επομένως, ο συνολικός αριθμός των παρατηρήσεων, των οποίων οι τιμές μόχλευσης τους υπερβαίνουν την τιμή $3*(p+1)/n$ είναι 2984.

Στην πορεία, αναπαριστούμε γραφικά τις αποστάσεις του Cook. Τοποθεντώντας το όρισμα id.n=4 μέσα στην συνάρτηση του plot δηλώνουμε πως θέλουμε να εμφανιστούν οι 4 μεγαλύτερες αποστάσεις του Cook.

```
> #Cook's distance
> plot(logit_model,which = 4, id.n=4)
> cooks.distance(logit_model)
> cooks.distance(logit_model)
           1            2            3            4            5
       6            7            8
5.169259e-07 1.908046e-06 3.969160e-07 5.937489e-07 1.068286e-06 1.213824
   e-06 1.213824e-06 1.603412e-04
.
.
.
          993          994          995          996          997
     998          999         1000
4.853351e-07 5.143108e-07 4.404973e-07 1.774072e-06 3.294798e-05 2.536594
   e-07 1.150182e-06 5.274650e-07
 [ reached getOption("max.print") -- omitted 71519 entries ]
> which(dfbeta(logit_model)>1)
integer(0)
```

Οι τιμές των αποστάσεων του Cook που επιστρέφονται είναι κατά πολύ μικρότερες της μονάδας, αυτό φαίνεται και από το Σχήμα 4.6. Επομένως θεωρούμε ότι καμία παρατήρηση δεν έχει έντονη επιρροή.

Ελέγχοντας επιπλέον τις τιμές DFBETA, για την επίδραση που έχει κάθε παρατήρηση στην εκτίμηση των συντελεστών, για το μοντέλο και συγκεκριμένα ποιές από τις παρατηρήσεις έχουν τιμή μεγαλύτερη της μονάδας, παρατηρούμε πως καμία παρατήρηση δεν ξεπερνάει την μονάδα. Οπότε, δεν υπάρχει λόγος ανησυχίας. Αντιλαμβανόμαστε ότι το μοντέλο δεν επηρεάζεται από μεμονωμένες παρατηρήσεις.



Σχήμα 4.6: Διάγραμμα αποστάσεων

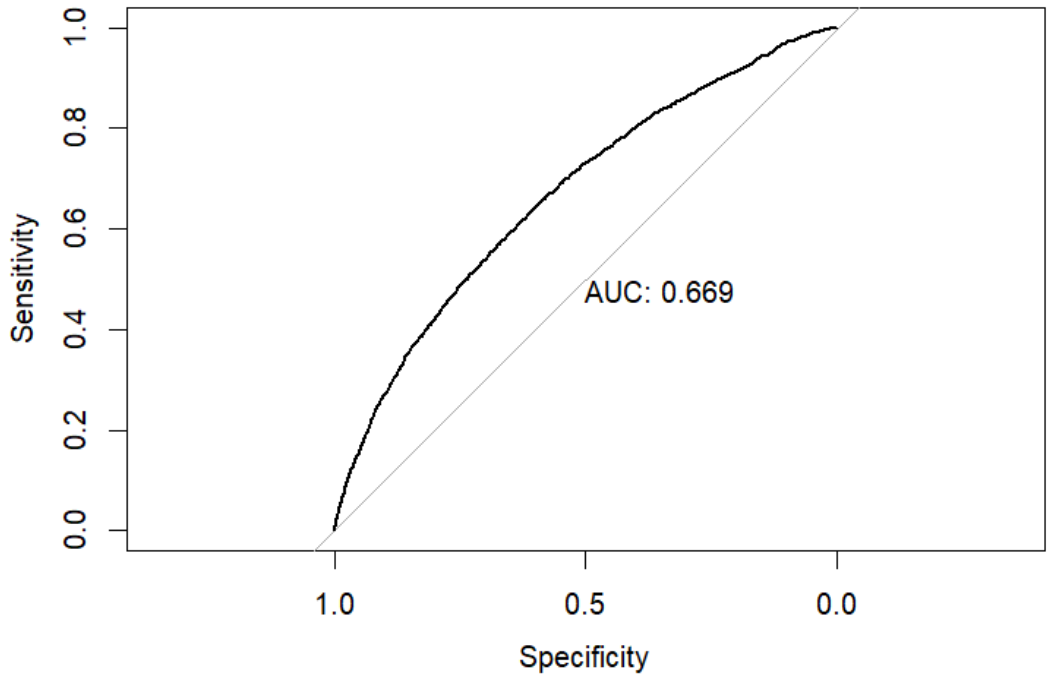
4.8 Ταξινόμηση

Μέσω της ταξινόμησης, οδηγούμαστε σε εντός δείγματος πρόβλεψη. Αρχικά, δημιουργούμε την καμπύλη ROC, στην οποία αναγράφεται η τιμή AUC. Στην συνέχεια, υπολογίζουμε το σημείο διαχωρισμού, $p*$, που μεγιστοποιεί τον δείκτη *Youden*.

```
#Classification
library(pROC)

DT<-data.table(age,gender,employment,education,citizenship)

model_glm<-glm(employment~gender+age+education+citizenship,data=DT,family
    = "binomial")

prob<-predict(model_glm, newdata=DT,type="response")

myroc<-roc(DT$employment~prob,plot=TRUE,print.auc=TRUE, color="blue")
```

```
> coords(myroc,"best")
  threshold specificity sensitivity
1 0.1209365    0.635259   0.6100118
```

Στο Σχήμα 4.7 παρατηρούμε ότι η τιμή $AUC = 0.669$. Άρα με πιθανότητα περίπου 70% το μοντέλο έχει την ικανότητα να διαχωρίσει την θετική από την αρνητική κλάση, δηλαδή τις τιμές $Y = 1, Y = 0$. Επομένως, καταλήγουμε στο συμπέρασμα ότι ο ταξινομητής μας είναι ικανοποιητικός.



Σχήμα 4.7: Η καμπύλη ROC της ταξινόμησης για το μοντέλο της λογιστής παλινδρόμησης

Κάντοντας χρήση του σημείου διαχωρισμού, $p* = 0.120936$, που υπολογίσαμε με την εντολή *coords*(*myroc*,"*best*") θα δημιουργήσουμε τον πίνακα σύγχυσης, *confusion matrix*. Με την χρήση του συγκεκριμένου πίνακα θα υπολογίσουμε τους τρείς βασικές τιμές: ακρίβεια, ευαισθησία και ειδικότητα.

```
>prediction<-ifelse(prob >=0.1209365, 1,0)
>prediction<-factor(prediction,levels=c(0:1))
>confusion_matrix<-table(prediction,DT$employment)
> confusion_matrix

prediction     1     0
         0 40675  3311
         1 23354  5179
```

```
accuracy<-sum(diag(confusion_matrix))/sum(confusion_matrix)
> accuracy
[1] 0.6323033

sensitivity<-confusion_matrix[2,2]/(confusion_matrix[2,2]+
    confusion_matrix[1,2])
> sensitivity
[1] 0.6100118

specificity<-confusion_matrix[1,1]/(confusion_matrix[1,1]+
    confusion_matrix[2,1])
> specificity
[1] 0.635259

youden<-specificity+sensitivity-1
> youden
[1] 0.2452708
```

Η ακρίβεια ισούται με 0.63, η ευαισθησία με 0.61, η ειδικότητα με 0.63 και ο δείκτης Youden με 0.24. Καταλήγουμε, και πάλι, ότι ο ταξινομητής είναι ικανοποιητικός.

4.9 Μέθοδος Διασταυρωμένης Επικύρωσης

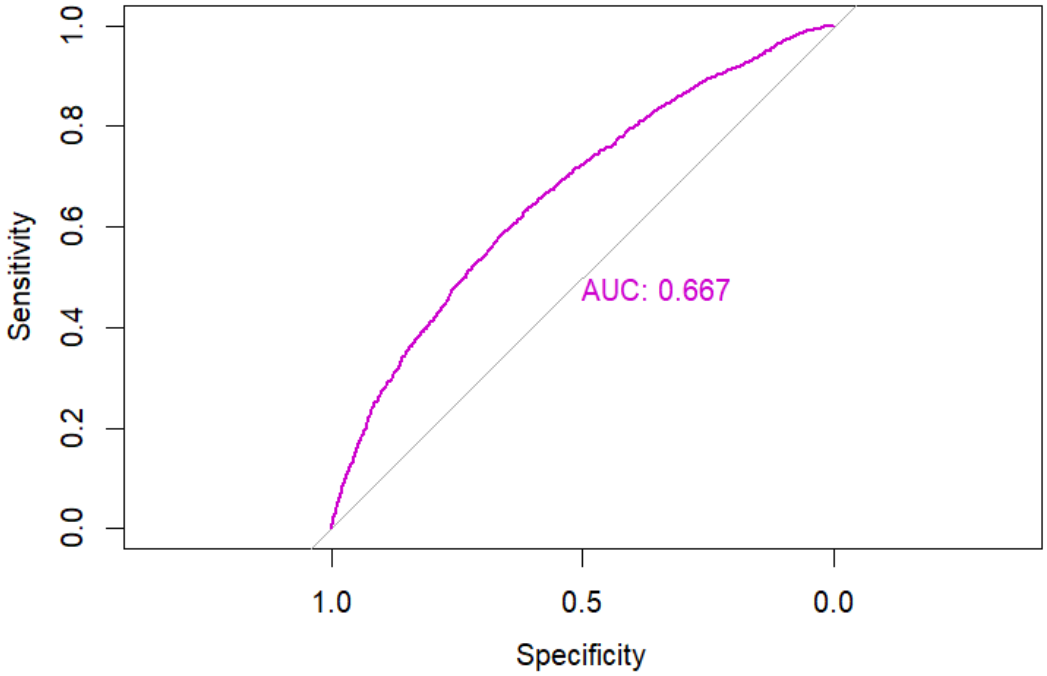
Για την μέθοδο της διασταυρωμένης επικύρωσης θα χρησιμοποιήσουμε τις βιβλιοθήκες *caret* και *e*1071. Η βιβλιοθήκη *caret* χρησιμοποιείται για την κατασκευή και αξιολόγηση μοντέλων μηχανικής μάθησης, συνδυαστικά με τους αλγορίθμους της *e*1071. Οι δύο αυτές βιβλιοθήκες χρησιμοποιούνται μαζί γιατί το *e*1071 παρέχει εξειδικευμένα εργαλεία για την ανάλυση της *caret*.

Για την υλοποίηση της μεθόδου, αρχικά, χωρίζουμε το σύνολο των δεδομένων μας σε σύνολο εκπαίδευσης και εξέτασης. Το σύνολο εκπαίδευσης, *train_data*, θα αποτελείται από το 70% των συνολικών παρατηρήσεων, και το σύνολο εξέτασης, *test_data*, θα είναι οι υπόλοιπες παρατηρήσεις.

```
library(caret)
library(e1071)
trainIndex <- createDataPartition(DT$employment, p = 0.7, list = FALSE)
train_data <- DT[trainIndex, ]
test_data <- DT[-trainIndex, ]
```

Στην συνέχεια, θα προσαρμόσουμε το μοντέλο μας στο σύνολο εκπαίδευσης και θα υπολογίζουμε τις προβλεπόμενες τιμές του μοντέλου πάνω στο σύνολο εξέτασης. Το είδος της πρόβλεψης, type="response", επιστρέφει τις πιθανότητες για την μεταβλητή απόκρισης. Τέλος, για την συγκεκριμένη μέθοδο κατασκευάζουμε την καμπύλη *ROC* του μοντέλου.

```
library(pROC)
model <- glm(employment~gender+age+education+citizenship, data =
    train_data, family = "binomial")
test_prob <- predict(model, test_data, type = "response")
test_roc=roc(test_data$employment~test_prob,plot=TRUE, print.auc=TRUE,col
    ="magenta3")
```



Σχήμα 4.8: Η καμπύλη ROC για το μοντέλο λογιστικής παλινδρόμησης με την μέθοδο διασταυρωμένης επικύρωσης

Στο Σχήμα 4.8, η *AUC* ισούται με 0.667, οπότε θεωρούμε τον εκτιμητή μας ικανοποιητικό.

Με τον παραπάνω τρόπο, έχουμε διαμερίσει το σύνολο δεδομένων μας μόνο μια φορά. Επομένως, τα αποτελέσματα μπορεί να έχουν υψηλή μεροληψία αν δεν έχουμε πλήρως αντιπροσωπευτικά δεδομένα ή υψηλή διασπορά ανάλογα με τον τρόπο που έχουν χωριστεί τα δεδομένα. Για την αντιμετώπιση αυτών των προβλημάτων θα κάνουμε χρήση της μεθόδου $10-fold$. Διαμερίζοντας το σύνολο μας σε δέκα διαφορετικά, ισοπληθή, σύνολα πετυχαίνουμε όλα τα δεδομένα μας να έχουν χρησιμοποιηθεί και σαν σύνολο εκπαίδευσης και σαν εξέτασης. Επομένως, τα αποτελέσματα μας είναι πιο ακριβή και για την χρήση του μοντέλου με άγνωστα δεδομένα.

Αρχικά, με την συνάρτηση *cut* δημιουργούμε δέκα ισοπληθή σύνολα δεδομένων. Στην συνέχεια, με την χρήση του βρόγχου *for* προσαρμόζουμε

το μοντέλο μας κάθε φορά σε ένα διαφορετικό σύνολο από τα δέκα που έχουμε δημιουργήσει και υπολογίζουμε πάλι τις τιμές των προβλεπόμενων πιθανοτήτων για το κάθε σύνολο εξέτασης. Τέλος, υπολογίζουμε τις τιμές της *AUC* για κάθε μια από τις δέκα φορές.

```
> library(ROCR)
> mydata<-DT[sample(nrow(DT)),]
> mydata<-data.frame(mydata,row.names = 1:72519)

> folds<-cut(seq(1,nrow(mydata)),breaks=10,labels=FALSE)
> auc_values<-numeric()

> for(i in 1:10) {
    +   testIndexes1<-which(folds==i,arr.ind = TRUE)
    +   testData<-mydata[testIndexes1,]
    +   Train_Data<-mydata[-testIndexes1,]
    +   model<-glm(employment~gender+age+education+citizenship,data=
   Train_Data,family="binomial")
    +   prob<-predict(model,testData,type="response")
    +   pred<-prediction(prob,testData$employment)
    +   auc<-performance(pred,measure="auc")
    +   auc_values[i]<-auc@y.values[[1]]
    + }

> mean(auc_values)
[1] 0.6316763
```

Η τιμή της *AUC* που προκύπτει σαν μέσος όρος των παραπάνω δέκα φορών είναι 0.632, άρα ο ταξινομητής μας είναι σχετικά ικανοποιητικός. Έχει την ικανότητα να διαχωρίζει τις κλάσεις $Y = 1$ από τις $Y = 0$ με 63,2% επιτυχία.

Χρησιμοποιούμε την παραπάνω διαδικασία και για τον υπολογισμό του βέλτιστου σημείου διαχωρισμού, $p*$, καθώς και των δεικτών της ακρίβειας, ευαισθησίας, ειδικότητας και του Youden, για αυτό το λόγο υπολογίζουμε και το δισδιάστο πίνακα σύγχυσης. Με την συνάρτηση *apply*() λαμβάνουμε τις μέσες τιμές των δεικτών από τις δέκα διαφορετικές φορές που έχουμε υπολογίσει. Υπενθυμίζουμε ότι το σημείου διαχωρισμού λαμβάνει τιμές στο $[0.01, 0, 99]$. Αν οι προβλεπόμενες πιθανότητες που προκύπτουν είναι μεγαλύτερες του σημείου διαχωρισμού θεωρούμε την μεταβλητή απόκρισης ως, $Y = 1$, διαφορετικά είναι ίση με 0.

Χωρίζουμε πάλι το σύνολο σε δέκα ισοπληθή σύνολα, όπου κάθε φορά χρησιμοποιείται ένα διαφορετικό ως σύνολο εξέτασης και προσαρμόζουμε το μοντέλο μας στο σύνολο εκπαίδευσης.

```
> library(ROCR)
> mydata<-mydata[sample(nrow(mydata)),]
> mydata<-data.frame(mydata,row.names = 1:72519)
> folds<-cut(seq(1,nrow(mydata)),breaks=10,labels=FALSE)
> cutoffs<-seq(0.01,0.99,0.01)
```

```
> accuracy<-matrix(nrow=10,ncol=length(cutoffs))
> sensitivity<-matrix(nrow=10,ncol=length(cutoffs))
> specificity<-matrix(nrow=10,ncol=length(cutoffs))
> youden<-matrix(nrow=10,ncol=length(cutoffs))


> for(j in 1:10){
    +   k<-0
    +   testIndexes<-which(folds==j,arr.ind = TRUE)
    +   testData<-mydata[testIndexes,]
    +   trainData<-mydata[-testIndexes,]
    +   model11<-glm(employment~gender+age+education+citizenship,data=
    trainData,family="binomial")
    +   prob<-predict(model11,newdata=testData,type="response")
    +   for(i in seq(along=cutoffs)){
    +     k<-k+1
    +     prediction<-ifelse(prob>=cutoffs[i],1,0)
    +     prediction<-factor(prediction,levels = c(0:1))
    +     confusion_matrix2<-table(prediction,testData$employment)
    +     accuracy[j,k]<-sum(diag(confusion_matrix2))/sum(
    confusion_matrix2)
    +     sensitivity[j,k]<-confusion_matrix2[2,2]/(confusion_matrix2
    [2,2]+confusion_matrix2[1,2])
    +     specificity[j,k]<-confusion_matrix2[1,1]/(confusion_matrix2
    [1,1]+confusion_matrix2[2,1])
    +     youden[j,k]<-specificity[j,k]+sensitivity[j,k]-1
    +
    +   }
    + }

> accuracy<-apply(accuracy,2,mean)
> sensitivity<-apply(sensitivity,2,mean)
> specificity<-apply(specificity,2,mean)
> youden<-apply(youden,2,mean)
> cutoffs[which.max(youden)]
[1] 0.12
> cutoffs[which.max(accuracy)]
[1] 0.52
```

Προκύπτουν δύο τιμές για το όριο p^*, ως μέγιστες τιμές του δείκτη Youden.

Χρησιμοποιώντας ως βέλτιστο σημείο διαχωρισμού την μέγιστη τιμή του δείτη Youden, 0.12 καταλήγουμε στα εξής αποτελέσματα.

```

> textIndexes<-48346:72519
> testData<-mydata[testIndexes,]
> trainData<-mydata[-testIndexes,]
> model_glm

Call:  glm(formula = employment ~ gender + age + education + citizenship,
           family = "binomial", data = DT)

Coefficients:
  (Intercept)       gender1           age    education1    education2
    education3    education4
```

```
0.19955       0.68426      -0.03649      -0.59459      -0.58059
   -0.82109      -0.91973
education5    education6  citizenship1  citizenship2
-1.44807      -1.96276       0.97506       0.43768

Degrees of Freedom: 72518 Total (i.e. Null);  72508 Residual
Null Deviance:      52370
Residual Deviance: 49540  AIC: 49560

> prob=predict(model_glm,newdata = testData,type="response")
> prediction<-ifelse(prob>=0.12,1,0)
> prediciton<-factor(prediction,levels = c(0:1))
> confusion_matrix<-table(prediction,testData$employment)
> confusion_matrix

prediction    1    0
           0 3968  341
           1 2390  553

> accuracy<-sum(diag(confusion_matrix))/sum(confusion_matrix)
> accuracy
[1] 0.6234142

> sensitivity<-confusion_matrix[2,2]/(confusion_matrix[2,2]+
   confusion_matrix[1,2])
> sensitivity
[1] 0.6185682

> specificity<-confusion_matrix[1,1]/(confusion_matrix[1,1]+
   confusion_matrix[2,1])
> specificity
[1] 0.6240956

> youden<-specificity+sensitivity-1
> youden
[1] 0.2426639
```

Παρατηρώντας τα παραπάνω αποτελέσματα η ακρίβεια προκύπτει ίση με 0.625, άρα η συνολική πιθανότητα να προβλέπει σωστά είναι 62,5%. Η ευαισθησία ίση με 0.619, άρα με πιθανότητα 61,9% προβλέπει σωστά τα εργαζόμενα άτομα. Η ειδικότητα ίση με 0.624, άρα με πιθανότητα 62,4% προβλέπει σωστά τους ανέργους και ο δείκτης Youden ίσος με 0.242, αρκετά χαμηλός, άρα η ισορροπία μεταξύ ευαισθησίας και ειδικότητας δεν είναι καλή. Στο σύνολο όμως ο ταξινομητής κρίνεται σχετικά ικανοποιητικός.

Κάνοντας χρήση του σημείου 0.12, θα υπολογίσουμε τον μέσο όρο της ακρίβειας του ταξινομητή, ακολουθώντας την ίδια διαδικασία.

```
> mydata<-mydata[sample(nrow(mydata)),]
> mydata<-data.frame(mydata,row.names = 1:72519)
> folds<-cut(seq(1,nrow(mydata)),breaks=10, labels = FALSE)
> accuracy<-NULL

```

```
  > for(j in 1:10){
    +   testIndexes<-which(folds==j,arr.ind = TRUE)
    +   testData<-mydata[testIndexes,]
    +   trainData<-mydata[-testIndexes,]
    +   model<-glm(employment~gender+age+education+citizenship,data=
    trainData,family="binomial")
    +   prob<-predict(model_glm,newdata = testData,type="response")
    +   prediction<-ifelse(prob>=0.12,1,0)
    +   prediciton<-factor(prediction,levels = c(0:1))
    +   confusion_matrix<-table(prediction,testData$employment)
    +
    +
    +   accuracy[j]<-sum(diag(confusion_matrix))/sum(confusion_matrix)
    +
    + }

> mean(accuracy)
[1] 0.6265254
```

Η μέση τιμή της ακρίβειας προκύπτει ίση με 0.627, άρα κατά πιθανότητα 62,7% προβλέπει σωστά. Επομένως, ο ταξινομητής μας είναι σχετικά ικανοποιητικός.

Θα επαναλάβουμε την παραπάνω διαδικασία χρησιμοποιώντας το 0.52 (την μέγιστη τιμή της ακρίβειας) ως βέλτιστο σημείο διαχωρισμού, $p*$.

```
> textIndexes<-48346:72519
> testData<-mydata[testIndexes,]
> trainData<-mydata[-testIndexes,]
>
  > prob=predict(model,newdata = testData,type="response")
> prediction<-ifelse(prob>=0.52,1,0)
> prediciton<-factor(prediction,levels = c(0:1))
> confusion_matrix<-table(prediction,testData$employment)
> confusion_matrix

prediction    1    0
0 6447  804
1    0    1
>
  > accuracy<-sum(diag(confusion_matrix))/sum(confusion_matrix)
> accuracy
[1] 0.889134
> sensitivity<-confusion_matrix[2,2]/(confusion_matrix[2,2]+
    confusion_matrix[1,2])
> sensitivity
[1] 0.001242236
> specificity<-confusion_matrix[1,1]/(confusion_matrix[1,1]+
    confusion_matrix[2,1])
> specificity
[1] 1
> youden<-specificity+sensitivity-1
> youden
[1] 0.001242236

```

Παρατηρώντας τα παραπάνω αποτελέσματα η ακρίβεια προκύπτει ίση με 0.889, άρα με πιθανότητα 88,9% προβλέπει σωστά, τιμή ιδιαίτερα αυξημένη. Η ευαισθησία ίση με 0.0012, άρα με πιθανότητα 0,12% προβλέπει τους εργαζομένους, αρκετά χαμηλή τιμή. Η ειδικότητα ίση με 1, άρα με πιθανότητα 100% προβλέπει τους ανέργους και ο δείκτης Youden ίσος με 0.0012, ακόμα πιο χαμηλός.

```
> mydata<-mydata[sample(nrow(mydata)),]
> mydata<-data.frame(mydata,row.names = 1:72519)
> folds<-cut(seq(1,nrow(mydata)),breaks=10, labels = FALSE)
> accuracy<-NULL
>
> for(j in 1:10){
+   testIndexes<-which(folds==j,arr.ind = TRUE)
+   testData<-mydata[testIndexes,]
+   trainData<-mydata[-testIndexes,]
+   model<-glm(employment~gender+age+education+citizenship,data=trainData
    ,family="binomial")
+   prob<-predict(model,newdata = testData,type="response")
+   prediction<-ifelse(prob>=0.52,1,0)
+   prediciton<-factor(prediction,levels = c(0:1))
+   confusion_matrix<-table(prediction,testData$employment)
+
+
+   accuracy[j]<-sum(diag(confusion_matrix))/sum(confusion_matrix)
+
+ }
> mean(accuracy)
[1] 0.8829962
```

Κάνοντας χρήση του σημείου 0.52 ο μέσος όρος της ακρίβειας προκύπτει μεγαλύτερος και ίσος με 0.883, άρα με 88,2% πιθανότητα προβλέπει σωστά.

Συγκρίνοντας τα αποτελέσματα και των δύο σημείων p^* οι δείκτες του πρώτου προσφέρουν μια πιο ισορροπημένη προσσέγιση, ενώ του δεύτερου πιο υψηλά ποσοστά μόνο στην ακρίβεια και την ειδικότητα. Άρα, αν οι εφαρμογές μας έχουν να κάνουν με την εύρεση των εργαζομένων η δεύτερη ανάλυση δεν θα έδινε ικανοποιητικά αποτελέσματα, σε αντίθεση αν θέλαμε να εστιάσουμε στους ανέργους.

4.10 Συμπεράσματα

Το μοντέλο που προσαρμόσαμε ανταποκρίνεται σε ολές τις προϋποθέσεις, πολυσυγγραμμικότητα, γραμμικότητα, ανεξαρτησία. Τα αποτελέσματα του έδειξαν ότι η ηλικία έχει αρνητική επίδραση στον λόγο συμπληρωματικών πιθανοτήτων, ο οποίος μειώνεται για τα άτομα μεγαλύτερης ηλικίας. Υπάρχει μεγάλη διαφορά στην άυξηση του λόγου όταν πρόκειται

για γυναίκα σε σχέση με άνδρα. Όσον αφορά την εκπαίδευση, η αύξηση των επιπέδων οδηγεί σε μείωση του λόγου. Τα άτομα με υπηκοότητα εκτός της Ελληνικής έχουν μικρότερο λόγο σε σχέση με τα υπόλοιπα. Όλοι οι συντελεστές είναι στατιστικά σημαντικοί. Η εντός δείγματος πρόβλεψη μέσω της ταξινόμησης μας έδειξε ότι ο ταξινομητής μας είναι σχετικά ικανοποιητικός. Ενώ με την Διασταυρωμένη Επικύρωση, καταλήξαμε ότι το μοντέλο μας θα είναι ικανοποιητική προσαρμογή και σε μελλοντικά δεδομένα, αν επιλέξουμε το πρώτο σημείο διαχωρισμού, p^*.

Κεφάλαιο 5

Σύγκριση Μοντέλων

Στο Κεφάλαιο 3 μιλήσαμε για την λογιστική παλινδρόμηση που χρησιμοποιεί ως συνάρτηση σύνδεσης την λογαριθμική συνάρτηση των συμπληρωματικών πιθανοτήτων, logit. Στην θέση της συνάρτησης logit μπορούν να χρησιμοποιηθούν και άλλες συνδετικές συναρτήσεις, μια εκ αυτών είναι η probit. Τα μοντέλα probit χρησιμοποιούν ως συνάρτηση σύνδεσης την αντίστροφη συνάρτηση της κατανομής της τυποποιημένης Κανονικής κατανομής. Η τυποποιημένη Κανονική κατανομή έχει μέση τιμή 0 και διασπορά 1 και η συνάρτηση κατανομής της είναι η

$$\Phi(z) = \mathbb{P}[Z \leq z] = \int_{-\infty}^{z} \frac{1}{\sqrt{2\pi}} e^{\frac{-x^2}{2}}\, dx.$$

Έστω p η πιθανότητα επιτυχίας, δηλαδή η μέση τιμή της μεταβλητής απόκρισης, τότε η αντίστροφη συνάρτηση της τυποποιημένης κανονικής κατανομής συμβολίζεται με $\Phi^{-1}(p)$.

5.1 Σύγκριση των δύο Μοντέλων στην R

5.1.1 Προσαρμογή Μοντέλου

Αρχικά, προσαρμόζουμε το μοντέλο στα δεδομένα χρησιμοποιώντας την συνάρτηση *glm*, όπως και με το μοντέλο *logit*. Η διαφορά σε σχέση με το προηγούμενο μοντέλο είναι ότι στο όρισμα *family* χρησιμοποιούμε την *probit*. Με την εντολή *summary* λαμβάνουμε τα αποτελέσματα του μοντέλου.

```
1 > probit_model<-glm(employment~gender+age+education+citizenship,data=dt2,
    family = binomial(link = "probit"))
2 > summary(probit_model)
3
4 Call:
```

```
glm(formula = employment ~ gender + age + education + citizenship,
    family = binomial(link = "probit"), data = dt2)

Coefficients:
               Estimate Std. Error z value Pr(>|z|)
(Intercept)   0.0189327  0.0853937   0.222    0.825
gender1       0.3520288  0.0127120  27.693  < 2e-16 ***
age          -0.0190999  0.0005572 -34.277  < 2e-16 ***
education1   -0.3425901  0.0826072  -4.147 3.37e-05 ***
education2   -0.3370988  0.0827902  -4.072 4.67e-05 ***
education3   -0.4627055  0.0813667  -5.687 1.30e-08 ***
education4   -0.5118251  0.0830314  -6.164 7.08e-10 ***
education5   -0.7913247  0.0820070  -9.649  < 2e-16 ***
education6   -1.0412421  0.0870118 -11.967  < 2e-16 ***
citizenship1  0.5641802  0.0649502   8.686  < 2e-16 ***
citizenship2  0.2590161  0.0305993   8.465  < 2e-16 ***
---


(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 52366  on 72518  degrees of freedom
Residual deviance: 49566  on 72508  degrees of freedom
AIC: 49588

Number of Fisher Scoring iterations: 5
```

Προφανώς οι συντελεστές του μοντέλου που προκύπτουν διαφέρουν από τους αντίστοιχους του μοντέλου *logit*.

Η σημαντικότητα του κάθε συντελεστή παραμένει ίδια, με τον σταθερό όρο να μην είναι ούτε εδώ στατιστικά σημαντικός, αφού λόγω της υψηλής $P - value$ δεν μπορούμε να απορρίψουμε την μηδενική υπόθεση. Το επίπεδο σημαντικότητας παραμένει $a = 5\%$.

Η ερμηνεία των συντελεστών σε ένα μοντέλο *probit* δεν είναι το ίδιο απλή, όπως στο μοντέλο *logit* λόγω της μη γραμμικότητας. Παρόλα αυτά, όταν ένας συντελεστής b_j είναι θετικός σημαίνει ότι η αύξηση της αντίστοιχης ανεξάρτητης μεταβλητής X_j οδηγεί σε αύξηση της πιθανότητας η Y να λάβει την τιμή 1, δηλαδή για την περίπτωση ένα άτομο να βρεί εργασία. Ενώ όταν ο συντελεστής b_j είναι αρνητικός, τα αποτελέσματα είναι αντίθετα.

5.1.2 Έλεγχοι Καλής Προσαρμογής

Υπολογίζοντας τους τρείς συντελεστές προσδιορισμού, παρατηρούμε ότι οι τιμές τους είναι αρκετά με τους αντίστοιχους του μοντέλου *logit*.

```

> mylprobit_chi_sq<-probit_model$null.deviance-probit_model$deviance
> myprobit_chi_sq_df<-probit_model$df.null-probit_model$df.residual
```

```
> pchisq(mylprobit_chi_sq,myprobit_chi_sq_df,lower.tail = FALSE)
[1] 0

> R2_L<-mylprobit_chi_sq/probit_model$null.deviance
> R2_L
[1] 0.05347747
>
> R2_CS<-1-exp((-mylprobit_chi_sq)/nrow(dt2))
> R2_CS
[1] 0.03788017
>
> R2_N<--R2_CS/(1-exp(probit_model$null.deviance/nrow(dt2)))
> R2_N
[1] 0.03577788
```

Οι τιμές που προκύπτουν είναι εξίσου μικρές με του μοντέλου *logit*. Οπότε μικρό ποσοστό της διασποράς της μεταβλητής απόκρισης εξηγείται.

```
> myprobit_null<-glm(employment~1,data=dt2,family = "binomial")
> anova(myprobit_null,probit_model,test="Chisq")
Analysis of Deviance Table

Model 1: employment ~ 1
Model 2: employment ~ gender + age + education + citizenship
  Resid. Df Resid. Dev Df Deviance  Pr(>Chi)
1     72518      52366
2     72508      49566 10   2800.4 < 2.2e-16 ***
---
```

Στην συνέχεια, υπολογίζουμε τους δείκτες *AIC* και *BIC* του μοντέλου.

```
> AIC(probit_model)
[1] 49587.82
> BIC(probit_model)
[1] 49688.93
```

Παρατηρούμε υψηλές τιμές, όπως στο προηγούμενο μοντέλο. Υπολογίζουμε την διαφορά των δεικτών ανάμεσα στα δύο μοντέλα, παρατηρούμε ότι η διαφορά είναι θετική, άρα οι δείκτες του μοντέλου *probit* είναι μεγαλύτεροι.

```
> AIC(probit_model)-AIC(logit_model)
[1] 25.82348
> BIC(probit_model)-BIC(logit_model)
[1] 25.82348
```

Με βάση τους συγκεκριμένους δείκτες επιλέγουμε το μοντέλο με τις μικρότερες τιμές, δηλαδή το μοντέλο *logit*. Αυτό μας προσφέρει καλύτερη ισορροπία μεταξύ καλής προσαρμογής και πολυπλοκότητας του μοντέλου.

5.1.3 Ταξινόμηση

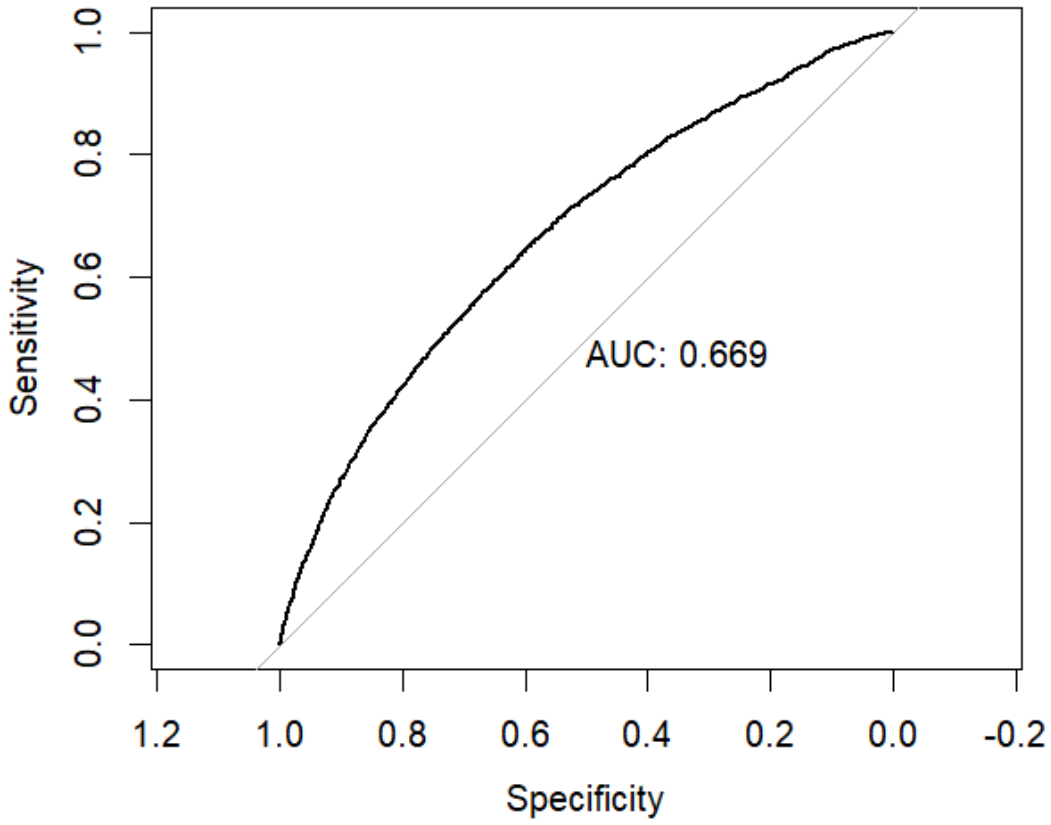
Πραγματοποιούμε και με αυτό το μοντέλο μια εντός δείγματος πρόβλεψη. Επομένως δημιουργούμε την καμπύλη ROC και υπολογίζουμε την τιμή AUC. Τέλος, έχοντας υπολογίσει το βέλτιστο σημείο διαχωρισμού, $p*$, δημιουργούμε τον πίνακα σύγχυσης και χρησιμοποιώντας τον υπολογίζουμε την ακρίβεια, την ευαισθησία, την ειδικότητα και τον Youden.

```
> library(pROC)
> mydata<-DT
> probit_model<-glm(employment~gender+age+education+citizenship,data=dt2,
    family = binomial(link = "probit"))
> prob1<-predict(probit_model, newdata=DT,type="response")
> roc_probit<-roc(DT$employment~prob1,plot=TRUE,print.auc=TRUE, color="
    red")
Setting levels: control = 1, case = 0
Setting direction: controls < cases
> coords(roc_probit,"best")
  threshold specificity sensitivity
1 0.1227467    0.631745   0.6135453
> prediction_probit<-ifelse(prob1 >=0.1227467, 1,0)
> prediction_probit<-factor(prediction_probit,levels=c(0:1))
> confusion_matrix_probit<-table(prediction_probit,DT$employment)
> confusion_matrix_probit

prediction_probit     1     0
                0 40450  3281
                1 23579  5209
> accuracy_probit<-sum(diag(confusion_matrix_probit))/sum(
    confusion_matrix_probit)
> accuracy_probit
[1] 0.6296143
> sensitivity_probit<-confusion_matrix_probit[2,2]/(
    confusion_matrix_probit[2,2]+confusion_matrix_probit[1,2])
> sensitivity_probit
[1] 0.6135453
> specificity_probit<-confusion_matrix_probit[1,1]/(
    confusion_matrix_probit[1,1]+confusion_matrix_probit[2,1])
> specificity_probit
[1] 0.631745
> youden_probit<-specificity_probit+sensitivity_probit-1
> youden_probit
[1] 0.2452903
```

Όπως παρατηρούμε στο Σχήμα 5.1 η καμπύλη ROC και η τιμή του AUC είναι ίδια με το μοντέλο *logit* και ίση με 0.669. Άρα, το μοντέλο με περίπου 70% μπορεί να διαχωρίσει την θετική από την αρνητική κλάση, όπως και το μοντέλο *logit*. Άρα, ο ταξινομητής που προκύπτει είναι ικανοποιητικός.

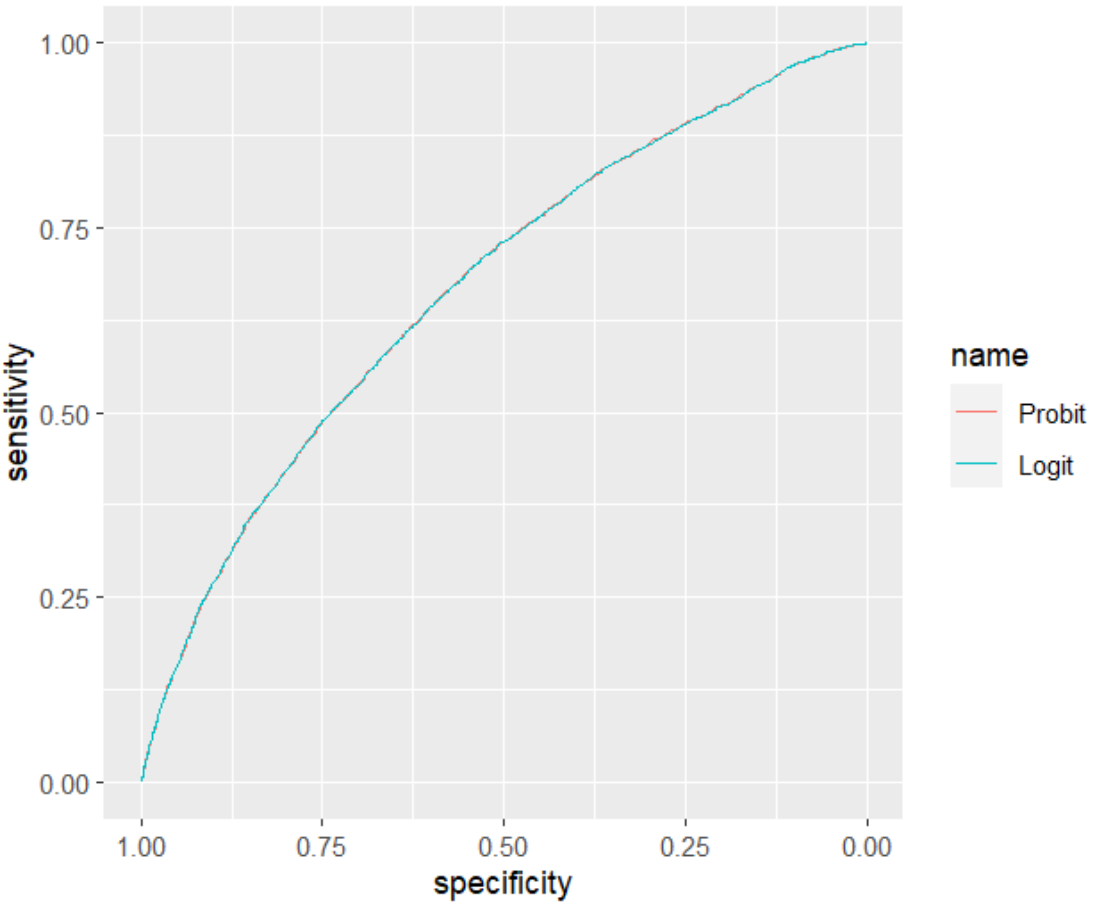
Στο Σχήμα 5.2 που απεικονίζονται οι καμπύλες *ROC* των δύο μοντέλων, βλέπουμε πως οι δύο καμπύλες σχεδόν ταυτίζονται.

```
ggroc(list(Probit = roc_probit, Logit = myroc))
```



Σχήμα 5.1: Η καμπύλη ROC για το μοντέλο Probit

Συμπεραίνουμε ότι τα δύο μοντέλα σχεδόν ταυτίζονται και έχουν παρόμοια προσαρμογή. Παρέχουν παρόμοιες προβλέψεις όσον αφορά στον υπολογισμό της πιθανότητας, $\mathbb{P}[Y = 1|X]$. Η μόνη διαφορά έγκειται όταν η πιθανότητα επιτυχίας είναι είτε πολύ μικρή (κοντά στο 0), είτε πολύ κοντά στο 1. Τότε, σύμφωνα με τον *Cox (1970)* η κανονική κατανομή προσεγγίζει το όριο πιο γρήγορα από την λογιστική. Αυτό συμβαίνει γιατί η κανονική κατανομή έχει λιγότερο έντονη ουρά στα άκρα, σε σχέση με την παχιά ουρά της λογιστικής κατανομής. Η λογιστική κατανομή είναι πιο ανθεκτική σε ακραίες τιμές. Η χρήση του μοντέλου *logit* είναι ευκολότερη και προσφέρει την εύκολη ερμηνεία των συντελεστών.



Σχήμα 5.2: Η καμπύλη ROC για τα μοντέλα Logit και Probit

Κεφάλαιο 6

Κώδικας Πινάκων-Σχημάτων Περιγραφικής Στατιστικής

6.1 Πακέτο data.table

Στο Κεφάλαιο 2, της Περιγραφικής Στατιστικής, παρουσιάζονται διάφοροι πίνακες που αποτελούνται από συχνότητες των μεταβλητών που χρησιμοποιούνται. Οι συγκεκριμένες συχνότητες έχουν υπολογιστεί με το πακέτο *data.table*. Το συγκεκριμένο πακέτο αποτελεί μια διαφορετική μορφή δομής δεδομένων, η οποία παρέχει ευκολία στον χειρισμό των δεδομένων, καθώς και στην γρηγορότερη εξαγωγή αποτελεσμάτων. Για τον χειρισμό των δεδομένων θεωρούμε οτι αποτελεί μια "τρισδιάστατη" δομή, η οποία καθορίζεται από τρείς δείκτες, *i* ο οποίος αναφέρεται στις γραμμές, *j* ο οποίος αναφέρεται στις στήλες και ο *by* με τον οποίο δημιουργούμε συναθροίσεις. Για τον υπολογισμό του συνολικού αριθμού παρατηρήσεων που ικανοποιούν μια συνθήκη χρησιμοποιούνται την εντολή .N. Επιπλέον, με την εντολή .(mean) μπορούμε να υπολογίσουμε την μέση τιμή για κάποια ποσοτική μεταβλητή που ικανοποιεί την συνθήκη.

Τέλος, η χρήση .() συμβολίζει την δημιουργία μιας νέας στήλης, ενώ η χρήση της μέσα στο *by* υποδηλώνει την ομαδοποίηση δύο μεταβλητών και πάνω.

Για να το χρησιμοποιήσουμε καλούμε το ομώνυμο πακέτο της R και μετατρέπουμε την υπάρχουσα μορφή δεδομένων σε data.table, με την εντολή *setDT*().

```
1 library(data.table)
2 df<-setDT(df)
3 dfe<-setDT(dfe)
4 dfu<-setDT(dfu)
```

```
Gender_Education<-df[, .N, by= .(employment, gender, education)]
Gender_Urbanisation <-df[, .N, by= .(employment, gender, urbanisation)]
```

```
Klados<-dfe[, .N, by=sector]
Symvasi_Urbanisation<- dfe[, .N, by= .(urbanisation, symvasi)]
Job<-dfe[, .N, by= .(job)]
jobU<-dfe[, .N, by=.(gender,job)]
jobE<-dfe[, .N, by=.(education,job)]
PositionG<-dfe[, .N, by=.(gender, position)]
PositionC<-dfe[, .N, by=.(citizenship, position)
```

```
mean_age_gender<-df[, .(mean_age = mean(age, na.rm=TRUE)), by=.(gender,
    employment)]
```

```
employment_gender_education <-df[, .N, by=.(employment, gender, education
    )]

employment_gender_urbanisation <-df[, .N, by=.(employment, gender,
    urbanisation)]

temployment_gender_citizenship <-df[, .N, by=.(employment, gender,
    citizenship)]

employment_gender_nuts <-df[, .N, by=.(employment, gender, nuts)]
```

```
hours_men <-df[gender == 1 & employment == 1, mean(hours, na.rm=TRUE)]

hours_women <-df[gender == 2 & employment == 1, mean(hours, na.rm =TRUE)]
```

```
hours_urbanisation1 <-df[employment == 1 & urbanisation == 1, mean(hours,
     na.rm=TRUE)]

hours_urbanisation2 <-df[employment == 1 & urbanisation == 2, mean(hours,
     na.rm=TRUE)]

hours_urbanisation3 <-df[employment == 1 & urbanisation == 3, mean(hours,
     na.rm=TRUE)]

hours_urbanisation4 <-df[employment == 1 & urbanisation == 4, mean(hours,
     na.rm=TRUE)]

hours_urbanisation5 <-df[employment == 1 & urbanisation == 5, mean(hours,
     na.rm=TRUE)]
```

```
Position<- df[, .N, by=.(gender, employment, position)]
```

```
Anazitisi<-dfu[, .N, by =anazitisi]
AnazitisiG<-dfu[, .N, by =anazitisi,gender]
AnazitisiE<-dfu[, .N, by =anazitisi,education]
```

```
Etos<- dfu[, .N, by =etos]
logos<-dfu[,.N,by=diakopi]
logosG<-dfu[,.N,by=diakopi,gender]
```

6.2 Πακέτο ggplot2

Όλα τα διαγράμματα της Περιγραφικής Στατιστικής του δευτέρου κεφαλαίου δημιουργήθηκαν με το πακέτο *ggplot2*. Σε αντίθεση με τις

υπάρχουσες συναρτήσεις της R για δημιουργία διαγραμμάτων, η βιβλιοθήκη *ggplot2* προσφέρει καλύτερα αποτελέσματα. Μας παρέχει την δυνατότητα προσαρμογής της μορφής των διαγραμμάτων. Τα γραφικά είναι καλύτερα, αφού μπορούμε να τροποποιήσουμε το χρώμα, το μέγεθος, να προσθέσουμε κείμενο, λεζάντες, καθώς και τίτλους στους άξονες. Η βιβλιοθήκη *ggplot2* παρέχει μια μεγάλη ποικιλία δυνατοτήτων, με την οποία μπορείς να παρέμβεις σε κάθε διάγραμμα. Μερικές από αυτές παρουσιάζονται παρακάτω.

Για να χρησιμοποιήσουμε το πακέτο καλούμε την ομώνυμη βιβλιοθήκη:

```
1 library(ggplot2)
```

Το βασικό πρότυπο που ακολουθείτε για την δημιουργία ενός διαγράμματος με την χρήση της συνάρτησης *ggplot* είναι:

```
1 ggplot(data=..., aes(...)) +  geom_function()
```

Αρχικά, καλούμε την συνάρτηση *ggplot* η οποία έχει τα εξής ορίσματα:

- *data =*, δηλώνει το σύνολο των δεδομένων που θα χρησιμοποιήσουμε για την απεικόνιση. Το όρισμα μπορεί να παραληφθεί και να μπεί στην θέση του το όνομα του συνόλου δεδομένων.
- Ακολουθεί το *aes*(), το οποίο αναφέρεται στα αισθητικά (aesthetics) του διαγράμματος. Με αυτό το όρισμα γίνεται η αντιστοίχιση των μεταβλητών που θα χρησιμοποιηθούν στον αντίστοιχο άξονα. Δηλαδή, *aes(x=..., y=...)*.

Μερικά από τα ορίσματα που χρησιμοποιούντα μέσα στο *aes*() είναι:

- Το color, το οποίο χρησιμοποιείται για τον χρωματισμό του διαγράμματος. Λαμβάνει ως τιμές κάποιο χρώμα από τις προεπιλογές της *R* ή κάποια κατηγορική μεταβλητή, άρα τα σημεία χρωματίζονται με διαφορετικά χρώματα ανάλογα με τις κατηγορίες της μεταβλητής.
- Το fill, χρησιμοποιείται για τον χρωματισμό του διαγράμματος όταν πρόκειται για ραβδοδιάγραμμα, ιστόγραμμα και θηκόγραμμα.
- Στην περίπτωση ραβδοδιαγράμματος, ιστογράμμτος και θηκογράμματος το color αναφέρεται στο χρώμα του περιγράμματος, ενώ το fill στο χρώμα με το οποίο γεμίζεται η κάθε μπάρα.
- Το alpha, αντιστοιχεί στο πόσο διαφανές ή πυκνό θέλουμε να είναι το χρώμα. Πεδίο ορισμού του είναι το $[0, 1]$, οι πιο χαμηλές τιμές αντιστοιχούν στα πιο διαφανή.

- Το size, ορίζει το μέγεθος των σημείων ενός διαγράμματος διασποράς.

Η συνάρτηση *geom_function()* χρησιμοποείται για να δηλώσουμε το είδος του διαγράμματος που θα χρησιμοποιήσουμε (διάγραμμα διασποράς, θηκογράμματα, ιστογράμματα κλπ). Στην παρούσα εργασία χρησιμοποιούνται τα εξής:

- geom_line: για την δημιουργία διαγραμμάτων στα οποία τα σημεία ενώνονται με γραμμές.
- geom_boxplot: για την δημιουργία θηκογράμματος.
- geom_histogram: για την δημιουργία ιστογράμματος.
- geom_bar: για την δημιουργία ραβδογράμματος.

Τα ορίσματα που αναφέρθηκαν παραπάνω για την *aes*() και αφορούσαν το χρώμα, το γέμισμα ή το μέγεθος μπορούν να χρησιμοποιηθούν και μέσα σε μια *geom_function()*.

Το όρισμα των ιστογραμμάτων *binwidth* χρησιμοποιείται για να δηλώσουμε το πλάτος των στηλών του.

Κάποια ορίσματα που χρησιμοποιούνται μέσα στην αντίστοιχη συνάρτηση *geom_function()* που έχουμε επιλέξει είναι:

- stat: αποτελεί μια παράμετρο των ραβδοδιαγραμμάτων και αναφέρεται στις τιμές του άξονα *y*. Οι τιμές που μπορεί να λάβει είναι η "identity", η οποία δηλώνει οτι οι τιμές θα εμφανίζονται με την ίδια σειρά που κατέχουν στο πλαίσιο δεδομένων.
- position: είναι ένα όρισμα που χρησιμοποιείται όταν σε ένα ραβδοδιάγραμμα έχει χρησιμοποιηθεί το όρισμα *fill* μέσα στο *aes*. Οι τιμές που μπορεί να λάβει είναι "stack", η οποία δημιουργεί στοιβαγμένο ραβδοδιάγραμμα όπου η κάθε μπάρα χωρίζεται σε διάφορα χρώματα και η "dodge", η οποία δημιουργεί ραβδοδιάγραμμα με τις μπάρες να βρίσκονται η μια δίπλα στην άλλη.

Η συνάρτηση *theme*() παρέχει την δυνατότητα επεξεργασίας των τίτλων ενός διαγράμματος, το μέγεθος, την θέση. Τα ορίσματα που χρησιμοποιούνται είναι:

- lengend.position: για τον ορισμό της θέσης του τίτλου (top, bottom, left, right, none).

- lengend.title: με το όρισμα element_text για τον ορισμό του μεγέθους του τίτλου.
- lengend.text: με το όρισμα element_text για τον ορισμό του μεγέθους του κειμένου.

Με τις εντολές:

- xlab, δίνουμε τίτλο στον άξονα x.
- ylab, δίνουμε τίτλο στον άξονα y.
- labs, δίνουμε τίτλο στο διάγραμμα (title), αλλάζουμε τον τίτλο του παραθέματος (color), προσθέτουμε περιγραφεί (caption).

Οι εντολές:

- scale_x _discrete(): ορίζει τις τιμές στον άξονα x.
- scale_y_discrete(): ορίζει τις τιμές στον άξονα y.

Η συνάρτηση *geom_text()* χρησιμοποείται για οτιδήποτε αφορά το κείμενο ενός διαγράμματος. Μερικά ορίσματα είναι:

- label,
- color, αναφέρεται στο χρώμα του κειμένου.
- size, αναφέρεται στο μέγεθος του κειμένου.
- v/hjust (verical/horizontal justification), ελέγχουν την κάθετη στοίχιση ενός κειμένου μέσα σε μια μπάρα, είτε ιστογράμματος, είτε ραβδοδιαγράμματος.

6.2.1 Διαγράμματα Κατάστασης Απασχόλησης

```
ggplot(Fylo, aes(x = group, y = value, fill = Fylo)) +
geom_bar(stat = "identity", position = "dodge") +
theme(legend.position = "top")
```

Listing 6.1: Διάγραμμα κατάστασης απασχόλησης ανα Φύλο

```
ggplot(ekpaideusi, aes(x = groupEd, y = valuesEd, fill =Ekpaideusi )) +
geom_bar(stat = "identity", position = "dodge") +
theme(legend.position = "top")+
theme(legend.title = element_text(size=10)) +
theme(legend.text = element_text(size=5))
```

Listing 6.2: Διάγραμμα κατάστασης απασχόλησης ανα Ανώτατο Επίπεδο Εκπαίδευσης

```
ggplot(ypikoothta, aes(x = groupCi, y = valuesCi, fill =Ypikoothta)) +
geom_bar(stat = "identity",
position = "dodge") +
```

```
theme(legend.position = "top")
```

Listing 6.3: Διάγραμμα κατάστασης απασχόλησης ανα Υπηκοότητα

```
df2<-df[!KATAP==3]
ggplot(df2, aes(x = as.factor(KATAP), y = AGE, fill = as.factor(KATAP)))
    +
  geom_boxplot(color="coral4", alpha=1.0) +
  scale_fill_manual(values = c("darkorchid1","coral" ))

hist_out<-ggplot(dfo, aes(x = ageo)) + geom_histogram(binwidth = 5, color
    ="darkred", fill="darkseagreen2")
hist_out
```

Listing 6.4: Διάγραμμα μέσης ηλικίας κατάστασης απασχόλησης

```
ggplot(astikopoihsh, aes(x = groupUr, y = valuesUr, fill =Astikopoihsh))
    +
geom_bar(stat = "identity",
position = "dodge") +
theme(legend.position = "top")+
theme(legend.title = element_text(size=5)) +
theme(legend.text = element_text(size=4))
```

Listing 6.5: Διάγραμμα κατάστασης απασχόλησης ανα Αστικοποίηση

```
ggplot(ypikoothta, aes(x = groupCi, y = valuesCi, fill =Ypikoothta)) +
geom_bar(stat = "identity",
position = "dodge") +
theme(legend.position = "top")
```

Listing 6.6: Διάγραμμα κατάστασης απασχόλησης ανα Υπηκοότητα

```
ggplot(perifereia, aes(x = groupNu, y = valuesNu, fill =Perifereia)) +
geom_bar(stat = "identity",
position = "dodge") +
theme(legend.position = "top")+
theme(legend.title = element_text(size=10)) +
theme(legend.text = element_text(size=5))
```

Listing 6.7: Διάγραμμα κατάστασης απασχόλησης ανα Περιφέρεια

6.2.2 Διαγράμματα Εργαζομένων

```
x <- ggplot(PartFull, aes(x = reorder(EidosApasxolisis, -Plithos), y =
    Plithos))
x <-x+  geom_bar(stat="identity",color="coral", fill="darkorchid1")+
geom_text(aes(label=Plithos),
vjust=1.6, color="white", size=3.5)
x
```

Listing 6.8: Διάγραμμα είδους απασχόλησης εργαζομένων

```
y <- ggplot(Thesi, aes(x = reorder(thesi, -Plithos), y = Plithos))
y <-y + geom_bar(stat="identity",color="coral", fill="darkorchid1")+
geom_text(aes(label=Plithos), vjust=1.6, color="white", size=3.5)
y
```

Listing 6.9: Διάγραμμα κύριας θέσης εργασίας εργαζομένων

```
z <- ggplot(pros_mon, aes(x = reorder(symvasi, -Plithos), y = Plithos))
z <-z+  geom_bar(stat="identity",color="coral", fill="darkorchid1")+
geom_text(aes(label=Plithos), vjust=1.6, color="white", size=3.5)
z
```

Listing 6.10: Διάγραμμα είδους σύμβασης εργαζομένων

```
h4<-hist(hours, breaks = "FD", plot = FALSE)
hist_hours_employed<-ggplot(df, aes(x = hours)) +
geom_histogram(binwidth = 5, color="coral", fill="darkorchid1")
hist_hours_employed
```

Listing 6.11: Ιστόγραμμα εβδομαδιαίων ωρών εργαζομένων

```

l <- ggplot(Oikonomikos, aes(x = reorder(klados, -Plithos), y = Plithos))
l <-l+  geom_bar(stat="identity",color="coral", fill="darkorchid1")+
geom_text(aes(label=Plithos), vjust=1.6, color="white", size=2.0)
l
```

Listing 6.12: Διάγραμμα κλάδου οικονομικής δραστηριότητας εργαζομένων

```
m <- ggplot(Ergasia, aes(x = reorder(thesi, -Plithos), y = Plithos))
m <-m+  geom_bar(stat="identity",color="coral", fill="darkorchid1")+
geom_text(aes(label=Plithos), vjust=1.6, color="white", size=3.0)
m
```

Listing 6.13: Διάγραμμα κύριας δραστηριότητας εργαζομένων

6.2.3 Διαγράμματα Ανέργων

```
x <- ggplot(Diakopi, aes(x = reorder(logos, -Plithos), y = Plithos))
x <-x+  geom_bar(stat="identity",color="darkorchid1", fill="coral")+
geom_text(aes(label=Plithos), vjust=1., color="white", size=3.)
+scale_x_discrete(limits=c("1","0","9","7","3","5","2","4","6","8"))
x
```

Listing 6.14: Διάγραμμα λόγου διακοπής τελευταίας εργασίας

```
u<- ggplot(Etos, aes(x = reorder(etos, -Plithos), y = Plithos))
u<-u+  geom_bar(stat="identity",color="darkorchid1", fill="coral")+
geom_text(aes(label=Plithos), vjust=1., color="white", size=1.5)
u
```

Listing 6.15: Διάγραμμα έτους τελευταίας εργασίας

```
k<- ggplot(Anazitisi, aes(x = reorder(diarkeia, -Plithos), y = Plithos))
k <-k+  geom_bar(stat="identity",color="darkorchid1", fill="coral")+
geom_text(aes(label=Plithos), vjust=1.6, color="white", size=3.5)
+scale_x_discrete(limits=c("< 6 mines","6-11 mines", "12 mines +"))
k
```

Listing 6.16: Διάγραμμα διάρκειας αναζήτησης εργασίας

```
y <- ggplot(Eggrafi, aes(x = reorder(eggrafi, -Plithos), y = Plithos))
y <-y+  geom_bar(stat="identity",color="darkorchid1", fill="coral")+
geom_text(aes(label=Plithos), vjust=1.3, color="white", size=3.5)
+scale_x_discrete(limits=c("Nai","Oxi"))
y
```

Listing 6.17: Διάγραμμα εγγραφής ως αναζητούσα/ων εργασίας

```
z <- ggplot(Epidoma, aes(x = reorder(epidoma, -Plithos), y = Plithos)))
z <-z+  geom_bar(stat="identity",color="darkorchid1", fill="coral")+
geom_text(aes(label=Plithos), vjust=1.6, color="white", size=3.5)
+scale_x_discrete(limits=c("Nai","Oxi"))
z
```

Listing 6.18: Διάγραμμα λήψης επιδόματος

6.2.4 Διαγράμματα Πορείας Ανεργίας, 2018-2022

```
ggplot(total_percentage, aes(x=year)) +
  geom_line(aes(y=total)) +
ylab("Pososto")+xlab("Etos")
```

Listing 6.19: Διάγραμμα εξέλιξης ανεργίας

```
ggplot(total_percentage, aes(x=year)) +
  geom_line(aes(y=man,color="Andres")) +
  geom_line(aes(y=woman, color="Gynaikes")) +
ylab("Pososto")+xlab("Etos")+labs(color = "Fylo")
```

Listing 6.20: Διάγραμμα εξέλιξης ανεργίας ανα Φύλο

```
ggplot(total_percentage, aes(x=year)) +
  geom_line(aes(y=Greece,color="Ellada")) +
  geom_line(aes(y=EU, color="E.E.")) +
  geom_line(aes(y=Other_Country, color="Alli Xwra"))+
ylab("Pososto")+xlab("Etos")+labs(color = "Ypikootita")
```

Listing 6.21: Διάγραμμα εξέλιξης ανα Υπηκοότητα

```
ggplot(total_percentage, aes(x=year)) +
  geom_line(aes(y=Attica,color="Perifereia Proteuousis")) +
  geom_line(aes(y=Thessaloniki, color="PS Thessalonikis")) +
  geom_line(aes(y=Urban, color="Astiki"))+
  geom_line(aes(y=Semiurban, color="Hmiastiki"))+
  geom_line(aes(y=Rural, color="Agrotiki"))
ylab("Pososto")+xlab("Etos")+labs(color = "Astikothta")
```

Listing 6.22: Διάγραμμα εξέλιξης ανα Αστικότητα

```
ggplot(total_percentage, aes(x=year)) +
  geom_line(aes(y=xoris_stoix,color="Xoris Stoixeiwdi Ekpaideusi")) +
  geom_line(aes(y=stoix, color="Stoixeiwdi Ekpaideusi")) +
  geom_line(aes(y=gymnasio, color="Gymnasio"))+
  geom_line(aes(y=lykeio, color="Lykeio"))+
  geom_line(aes(y=iek, color="IEK"))+
  geom_line(aes(y=proptyxiako, color="Proptyxiako"))+
  geom_line(aes(y=metaptyxiako, color="Metaptyxiako"))+
ylab("Pososto")+xlab("Etos")+labs(color = "Anwtato epipedo ekpaideusis")
```

Listing 6.23: Διάγραμμα εξέλιξης ανεργίας ανα Εκπαίδευση

```
ggplot(total_percentage, aes(x=year)) +
  geom_line(aes(y=age2, color="15-19")) +
  geom_line(aes(y=age3, color="20-24"))+
  geom_line(aes(y=age4, color="25-29"))+
  geom_line(aes(y=age5, color="30-34"))+
```

```
  geom_line(aes(y=age6, color="35-39"))+
  geom_line(aes(y=age7, color="40-44"))+
  geom_line(aes(y=age8,color="45-49")) +
  geom_line(aes(y=age9, color="50-54")) +
  geom_line(aes(y=age10, color="55-59"))+
  geom_line(aes(y=age11, color="60-64"))+
  geom_line(aes(y=age12, color="65-69"))+
  geom_line(aes(y=age13, color="70-74"))+
ylab("Pososto")+xlab("Etos")+labs(color = "5eteis ilikiakes omades")
```

Listing 6.24: Διάγραμμα εξέλιξης ανεργίας ανα 5ετείς ηλικιακές ομάδες

```
ggplot(total_percentage, aes(x=year)) +
  geom_line(aes(y=ana_Macedonia_Thrace,color="A. Macedonia & Thraki")) +
  geom_line(aes(y=central_Macedonia, color="Kentriki Macedonia")) +
  geom_line(aes(y=wes_Macedonia, color="Dytiki Macedonia"))+
  geom_line(aes(y=epirus, color="Hpeiros"))+
  geom_line(aes(y=thessaly, color="Thessalia"))+
  geom_line(aes(y=ionian, color="Ionia Nisia"))+
  geom_line(aes(y=wes_Greece, color="Dytiki Ellada"))+
  geom_line(aes(y=sterea,color="Sterea Ellada")) +
  geom_line(aes(y=Attica, color="Attiki")) +
  geom_line(aes(y=Peloponnese, color="Peloponnisos"))+
  geom_line(aes(y=N_Aegean, color="Boreio Aigaio"))+
  geom_line(aes(y=S_Aegean, color="Notio Aigaio"))+
  geom_line(aes(y=Crete, color="Kriti"))+
ylab("Pososto")+xlab("Etos")+labs(color = "Perifereia")
```

Listing 6.25: Διάγραμμα εξέλιξης ανεργίας ανα Περιφέρεια

Παράρτημα Α΄

Ακρωνύμια και Συντομογραφίες

PAPI Paper Assisted Personal Interviewing

CAPI Computer Assisted Personal Interviewing

ISCO International Standard Classification of Occupations

Η Διεθνής Πρότυπη Ταξινόμηση των Επαγγελμάτων αποτελεί δομή ταξινόμησης της Διεθνούς Οργάνωσης Εργασίας (International Labor Organisation, ILO) για την ταξινόμηση της πληροφορίας που σχετίζεται με την εργασίας και τις θέσεις εργασίας. Αποτελεί μέρος της οικονομικής και κοινωνικής ταξινόμησης των Ενωμένων Εθνών. Η τρέχουσα έκδοση υιοθετήθηκε το 2008 και είναι γνωστή ως ISC0-08.

NACE Rev. 2/ ΣΤΑΚΟΔ-08 Η ΣΤΑΚΟΔ-08 αποτελεί στατιστική ταξινόμηση των οικομομικών δραστηριοτήτων και βασίζεται στην Ταξινόμηση των Οικονομικών Δραστηριοτήτων NACE Rev. 2 της Ευρωπαϊκής Ένωσης.

ILO International Labor Organisation

Διεθνής Οργανισμός Εργασίας

Βιβλιογραφία

Ελληνική

[1] *Δημήτριος Φουσκάκης. (2021). Ανάλυση Δεδομένων με χρήση της R, εκδόσεις Τσότρας.*

[2] *Joel Grus (2021). Επιστήμη Δεδομένων: Βασικές Αρχές και Εφαρμογές με Python, εκδόσεις Παπασωτηρίου.*

Αγγλική

[3] *Raymod H. Myers, Douglas C. Montgomery, G. Geoffrey Vining and Timothy J. Robinson (2010). Generalized Linear Models with applications in Engineering and the Sciences, Wiley.*

[4] *Hsiao, C. (1996). Logit and Probit Models. In: Mátyás, L., Sevestre, P. (eds) The Econometrics of Panel Data. Advanced Studies in Theoretical and Applied Econometrics, vol 33. Springer, Dordrecht.*

[5] *Annette J. Dobson (2002). An Introduction to Generalized Linear Models, Chapman and Hall/CRC.*

[6] *Guo Chen and Hiroki Tsurumi (2010). Probit and Logit Model Selection, Communications in Statistics - Theory and Methods, 40:1, 159-175.*

Επιπλέον Βιβλιογραφία

[7] *Πηγή δεδομένων εργασίας, Αρχεία Δημόσιας Χρήσης ΕΛΣΤΑΤ.*

[8] *Συνοπτική έκθεση ποιότητας για χρήστες, ΕΛΣΤΑΤ.*

[9] *Ποσοστά ανεργίας στην Ευρώπη, Συνολικές εβδομαδιαίες ώρες εργασίας στην Ευρώπη: Βάση δεδομένων EUROSTAT για LABOR FORCE SURVEY.*

[10] *Sarang Narkhede (2018). Understanding AUC - ROC Curve.*